1 of 1 DOCUMENT

DANIEL J. CAPON, ARTHUR WEISS, BRIAN A. IRVING, MARGO R. ROBERTS, and KRISZTINA ZSEBO, Appellants, v. ZELIG ESHHAR, DANIEL SCHINDLER, TOVA WAKS, and GIDEON GROSS, Cross-Appellants, v. JON DUDAS, Director of the Patent and Trademark Office, Intervenor.

03-1480, 03-1481

UNITED STATES COURT OF APPEALS FOR THE FEDERAL CIRCUIT

2005 U.S. App. LEXIS 16865

August 12, 2005, Decided

PRIOR HISTORY: [*1] Appealed from: United States Patent and Trademark Office Board of Patent Appeals and Interferences. (Interference No. 103,887)

LexisNexis(R) Headnotes

COUNSEL: Steven B. Kelber, Piper Rudnick, LLP, of Washington, DC, argued for appellants.

Roger L. Browdy, Browdy and Neimark, P.L.L.C., of Washington, DC, argued for cross-appellants.

Mary L. Kelly, Associate Solicitor, Office of the Solicitor, United States Patent and Trademark Office, of Arlington, Virginia, argued for intervenor. With her on the brief were John M. Whealan, Solicitor and Stephen Walsh, Associate Solicitor.

JUDGES: Before NEWMAN, MAYER, * and GAJARSA, Circuit Judges.

* Haldane Robert Mayer vacated the position of Chief Judge on December 24, 2004.

OPINIONBY: NEWMAN

OPINION: NEWMAN, Circuit Judge.

Both of the parties to a patent interference proceeding have appealed the decision of the Board of Patent Appeals and Interferences of the United States Patent and Trademark Office, wherein the Board held that the specification of neither party met the written description requirement of the patent statute. Capon v. Eshhar, Interf. No. 103,887 (Bd. Pat. App. & Interf. Mar. 26, 2003). The Board dissolved the interference and cancelled all

[*2] of the claims of both parties corresponding to the interference count. With this ruling, the Board terminated the proceeding and did not reach the question of priority of invention. We conclude that the Board erred in its application of the law of written description. The decision is vacated and the case is remanded to the Board for further proceedings.

BACKGROUND

Daniel J. Capon, Arthur Weiss, Brian A. Irving, Margo R. Roberts, and Krisztina Zsebo (collectively "Capon") and Zelig Eshhar, Daniel Schindler, Tova Waks, and Gideon Gross (collectively "Eshhar") were the parties to an interference proceeding between Capon's *United States Patent No. 6,407,221* ("the *'221 patent*") entitled "Chimeric Chains for Receptor-Associated Signal Transduction Pathways" and Eshhar's patent application Serial No. 08/084,994 ("the '994 application") entitled "Chimeric Receptor Genes and Cells Transformed Therewith." Capon's *Patent No. 5,359,046* ("the '046 patent"), parent of the *'221 patent*, was also included in the interference but was held expired for non-payment of a maintenance fee. The PTO included the *'046* patent in its decision and in its argument of this appeal. n1

n1 Although Capon is designated as appellant and Eshhar as cross-appellant, both appealed the Board's decision. See *Fed. R. App. P. 28(h)*. The Director of the PTO intervened to support the Board, and has fully participated in this appeal.

[*3]

A patent interference is an administrative proceeding pursuant to *35 U.S.C. §§ 102(g)* and *135(a)*, conducted for the purpose of determining which of competing applicants is the first inventor of common subject matter. An interference is instituted after the separate patent applications have been examined and found to contain patentable subject matter. Capon's patents had been examined and had issued before this interference was instituted, and Eshhar's application had been examined and allowed but a patent had not yet issued.

During an interference proceeding the Board is authorized to determine not only priority of invention but also to redetermine patentability. *35 U.S.C. § 6(b)*. The question of patentability of the claims of both parties was raised *sua sponte* by an administrative patent judge during the preliminary proceedings. Thereafter the Board conducted an *inter partes* proceeding limited to this question, receiving evidence and argument. The Board then invalidated all of the claims that had been designated as corresponding to the count of the interference, viz., all of the claims of the Capon *'221 patent*, claims 5-8 of

[*4] the Capon *'046* patent, and claims 1-7, 9-20, and 23 of the Eshhar '994 application.

In accordance with the Administrative Procedure Act, the law as interpreted and applied by the agency receives plenary review on appeal, and the agency's factual findings are reviewed to determine whether they were arbitrary, capricious, or unsupported by substantial evidence in the administrative record. See *5 U.S.C. § 706(2)*; *Dickinson v. Zurko, 527 U.S. 150, 164-65, 144 L. Ed. 2d 143, 119 S. Ct. 1816 (1999)*; *In re Gartside, 203 F.3d 1305, 1315 (Fed. Cir. 2000)*.

The Invention

A chimeric gene is an artificial gene that combines segments of DNA in a way that does not occur in nature. The *'221 patent* and '994 application are directed to the production of chimeric genes designed to enhance the immune response by providing cells with specific cell-surface antibodies in a form that can penetrate diseased sites, such as solid tumors, that were not previously reachable. The parties explain that their invention is a way of endowing immune cells with antibody-type specificity, by combining known antigen-binding-domain producing DNA and known lymphocyte-receptor-protein

[*5] producing DNA into a unitary gene that can express a unitary polypeptide chain. Eshhar summarized the problem to which the invention is directed:

> Antigen-specific effector lymphocytes, such as tumor-specific T cells, are very rare, individual-specific, limited in their recognition spectrum and difficult to obtain against most malignancies. Antibodies, on the other hand, are readily obtainable, more easily derived, have wider spectrum and are not individual-specific. The major problem of applying specific antibodies for cancer immunotherapy lies in the inability of sufficient amounts of monoclonal antibodies (mAb) to reach large areas within solid tumors.

Technical Paper Explaining Eshhar's Invention, at 6.

The inventions of Capon and Eshhar are the chimeric DNA that encodes single-chain chimeric proteins for expression on the surface of cells of the immune system, plus expression vectors and cells transformed by the chimeric DNA. The experts for both parties explain that the invention combines selected DNA segments that are both endogenous and nonendogenous to a cell of the immune system, whereby the nonendogenous segment encodes the single-chain variable ("scFv")

[*6] domain of an antibody, and the endogenous segment encodes cytoplasmic, transmembrane, and extracellular domains of a lymphocyte signaling protein. They explain that the scFv domain combines the heavy and light variable ("Fv") domains of a natural antibody, and thus has the same specificity as a natural antibody. Linking this single chain domain to a lymphocyte signaling protein creates a chimeric scFv-receptor ("scFvR") gene which, upon transfection into a cell of the immune system, combines the specificity of an antibody with the tissue penetration, cytokine production, and target-cell destruction capability of a lymphocyte.

The parties point to the therapeutic potential if tumors can be infiltrated with specifically designed immune cells of appropriate anti-tumor specificity.

The Eshhar Claims

The Board held unpatentable the following claims of Eshhar's '994 application; these were all of the '994 claims that had been designated as corresponding to the count of the interference. Eshhar's claim 1 was the designated count.

1. A chimeric gene comprising

a first gene segment encoding a single-chain Fv domain (scFv) of a specific antibody and

a second gene segment

[*7] encoding partially or entirely the transmembrane and cytoplasmic, and optionally the extracellular, domains of an endogenous protein

wherein said endogenous protein is expressed on the surface of cells of the immune system and triggers activation and/or proliferation of said cells,

which chimeric gene, upon transfection to said cells of the immune system, expresses said scFv domain and said domains of said endogenous protein in one single chain on the surface of the transfected cells such that the transfected cells are triggered to activate and/or proliferate and have MHC nonrestricted antibody-type specificity when said expressed scFV domain binds to its antigen.

2. A chimeric gene according to claim 1 wherein the second gene segment further comprises partially or entirely the extracellular domain of said endogenous protein.

3. A chimeric gene according to claim 1 wherein the first gene segment encodes the scFv domain of an antibody against tumor cells.

4. A chimeric gene according to claim 1 wherein the first gene segment encodes the scFv domain of an antibody against virus infected cells.

5. A chimeric gene according to claim 4 wherein the virus is HIV.

6.

[*8] A chimeric gene according to claim 1 wherein the second gene segment encodes a lymphocyte receptor chain.

7. A chimeric gene according to claim 6 wherein the second gene segment encodes a chain of the T cell receptor.

9. A chimeric gene according to claim 7 wherein the second gene segment encodes the a, B, y, or o chain of the antigen-specific T cell receptor.

10. A chimeric gene according to claim 1 wherein the second gene segment encodes a polypeptide of the TCR/CD3 complex.

11. A chimeric gene according to claim 10 wherein the second gene segment encodes the zeta or eta isoform chain.

12. A chimeric gene according to claim 1 wherein the second gene segment encodes a subunit of the Fc receptor or IL-2 receptor.

13. A chimeric gene according to claim 12 wherein the second gene segment encodes a common subunit of IgE and IgG binding Fc receptors.

14. A chimeric gene according to claim 13 wherein said subunit is the gamma subunit.

15. A chimeric gene according to claim 13 wherein the second gene segment encodes the CD16a chain of the FcyRIII or FcyRII.

16. A chimeric gene according to claim 12 wherein the second gene segment encodes the a

[*9] or B subunit of the IL-2 receptor.

17. An expression vector comprising a chimeric gene according to claim 1.

18. A cell of the immune system endowed with antibody specificity transformed with an expression vector according to claim 17.

19. A cell of the immune system endowed with antibody specificity comprising a chimeric gene according to claim 1.

20. A cell if the immune system according to claim 19 selected from the group consisting of a natural killer cell, a lymphokine activated killer cell, a cytotoxic T cell, a helper T cell and a subtype thereof.

23. A chimeric gene according to claim 1 wherein said endogenous protein is a lymphocyte receptor chain, a polypeptide of the TCR/CD3 complex, or a subunit of the Fc or IL-2 receptor.

The Board did not discuss the claims separately, and held that the specification failed to satisfy the written description requirement as to all of these claims.

The Capon Claims

Claims 1-10, all of the claims of the *'221 patent*, were held unpatentable on written description grounds. Claims 1-6 are directed to the chimeric DNA, claims 7, 8, and 10 to the corresponding cell comprising the DNA, and claim 9 to

[*10] the chimeric protein:

1. A chimeric DNA encoding a membrane bound protein, said chimeric DNA comprising in reading frame:

DNA encoding a signal sequence which directs said membrane bound protein to the surface membrane;

DNA encoding a non-MHC restricted extracellular binding domain which is obtained from a single chain antibody that binds specifically to at least one ligand, wherein said at least one ligand is a protein on the surface of a cell or a viral protein;

DNA encoding a transmembrane domain which is obtained from a protein selected from the group consisting of CD4, CD8, immunoglobulin, the CD3 zeta chain, the CD3 gamma chain, the CD3 delta chain and the CD3 epsilon chain; and

DNA encoding a cytoplasmic signal-transducing domain of a protein that activates an intracellular messenger system which is obtained from CD3 zeta,

wherein said extracellular domain and said cytoplasmic domain are not naturally joined together, and said cytoplasmic domain is not naturally joined to an extracellular ligand-binding domain, and when said chimeric DNA is expressed as a membrane bound protein in a host cell under conditions suitable for expression, said membrane bound protein

[*11] initiates signaling in said host cell when said extracellular domain binds said at least one ligand.

2. The DNA of claim 1, wherein said single-chain antibody recognizes an antigen selected from the group consisting of viral antigens and tumor cell associated antigens.

3. The DNA of claim 2 wherein said single-chain antibody is specific for the HIV env glycoprotein.

4. The DNA of claim 1, wherein said transmembrane domain is naturally joined to said cytoplasmic domain.

5. An expression cassette comprising a transcriptional initiation region, the DNA of claim 1 under the transcriptional control of said transcriptional initiation region, and a transcriptional termination region.

6. A retroviral RNA or DNA construct comprising the expression cassette of claim 5.

7. A cell comprising the DNA of claim 1.

8. The cell of claim 7, wherein said cell is a human cell.

9. A chimeric protein comprising in the N-terminal to C-terminal direction:

a non-MHC restricted extracellular binding domain which is obtained from a single chain antibody that binds specifically to at least one ligand, wherein said at least one ligand is a protein on the surface of a cell

[*12] or a viral protein;

a transmembrane domain which is obtained from a protein selected from the group consisting CD4, CD8, immunoglobulin, the CD3 zeta chain, the CD3 gamma chain, the CD3 delta chain and the CD3 epsilon chain; and

a cytoplasmic signal-transducing domain of a protein that activates an intracellular messenger system which is obtained from CD3 zeta,

wherein said extracellular domain and said cytoplasmic domain are not naturally joined together, and said cytoplasmic domain is not naturally joined to an extracellular ligand-binding domain, and when said chimeric protein is expressed as a membrane bound protein in a host cell under conditions suitable for expression, said membrane bound protein initiates signaling in said host cell when said extracellular domain binds said at least one ligand.

10. A mammalian cell comprising as a surface membrane protein, the protein of claim 9.

In addition, claims 5, 6, 7, and 8 of Capon's *'046* patent were held unpatentable. These claims are directed to chimeric DNA sequences where the encoded extracellular domain is a single-chain antibody containing ligand binding activity.

The Board Decision

The Board presumed

[*13] enablement by the specifications of the *'221 patent* and '994 application of the full scope of their claims, and based its decision solely on the ground of failure of written description. The Board held that neither party's specification provides the requisite description of the full scope of the chimeric DNA or encoded proteins, by reference to knowledge in the art of the "structure, formula, chemical name, or physical properties" of the DNA or the proteins. In the Board's words:

> We are led by controlling precedent to understand that the full scope of novel chimeric DNA the parties claim is not described in their specifications under *35 U.S.C. § 112*, first paragraph, by reference to contemporary and/or prior knowledge in the art of the structure, formula, chemical name, or physical properties of many protein domains, and/or DNA sequences which encode many protein domains, which comprise single-chain proteins and/or DNA constructs made in accordance with the plans, schemes, and examples thereof the parties disclose.

Bd. op. at 4. As controlling precedent the Board cited *Regents of the University of California v. Eli Lilly & Co., 119 F.3d 1559 (Fed. Cir. 1997)*;

[*14] *Fiers v. Revel, 984 F.2d 1164 (Fed. Cir. 1993)*; *Amgen, Inc. v. Chugai Pharmaceutical Co., 927 F.2d 1200 (Fed. Cir. 1991)*; and *Enzo Biochem, Inc. v. Gen-Probe, Inc., 296 F.3d 1316 (Fed. Cir. 2002)*. The Board summarized its holding as follows:

> Here, both Eshhar and Capon claim novel genetic material described in terms of the functional characteristics of the protein it encodes. Their specifications do not satisfy the written description requirement because persons having ordinary skill in the art would not have been able to visualize and recognize the identity of the claimed genetic material without considering additional knowledge in the art, performing additional experimentation, and testing to confirm results.

Bd. op. at 89.

DISCUSSION

Eshhar and Capon challenge both the Board's interpretation of precedent and the Board's ruling that their descriptions are inadequate. Both parties explain that their chimeric genes are produced by selecting and combining known heavy-and light-chain immune-related DNA segments, using known DNA-linking procedures. The specifications of both parties describe procedures for identifying

[*15] and obtaining the desired immune-related DNA segments and linking them into the desired chimeric genes. Both parties point to their specific examples of chimeric DNA prepared using identified known procedures, along with citation to the scientific literature as to every step of the preparative method.

The parties presented expert witnesses who placed the invention in the context of prior knowledge and explained how the descriptive text would be understood by persons of skill in the field of the invention. The witnesses explained that the principle of forming chimeric genes from selected segments of DNA was known, as well as their methods of identifying, selecting, and combining the desired segments of DNA. Dr. Eshhar presented an expert statement wherein he explained that the prior art contains extensive knowledge of the nucleotide structure of the various immune-related segments of DNA; he stated that over 785 mouse antibody DNA light chains and 1,327 mouse antibody DNA heavy chains were known and published as early as 1991. Similarly Capon's expert Dr. Desiderio discussed the prior art, also citing scientific literature:

> The linker sequences disclosed in the *'221 patent*

> [*16] (col. 24, lines 4 and 43) used to artificially join a heavy and light chain nucleic acid sequence and permit functional association of the two ligand binding regions were published by 1990, as were the methods for obtaining the mature sequences of the desired heavy and light chains for constructing a SAb (Exhibit 47, Batra et al., J., Biol. Chem., 1990; Exhibit 48, Bird et al., Science, 1988; Exhibit 50, Huston et al., PNAS, 1988; Exhibit 51, Chaudhary, PNAS, 1990, Exhibit 56, Morrison et al., Science, 1985; Exhibit 53, Sharon et al., Nature 1984).

Desiderio declaration at 4 P11.

Both parties stated that persons experienced in this field would readily know the structure of a chimeric gene made of a first segment of DNA encoding the single-chain variable region of an antibody, and a second segment of DNA encoding an endogenous protein. They testified that re-analysis to confirm these structures would not be needed in order to know the DNA structure of the chimeric gene, and that the Board's requirement that the specification must reproduce the "structure, formula, chemical name, or physical properties" of these DNA combinations had been overtaken by the state of the science.

[*17] They stated that where the structure and properties of the DNA components were known, reanalysis was not required.

Eshhar's specification contains the nucleotide sequences of sixteen different receptor primers and four different scFv primers from which chimeric genes encoding scFvR may be obtained, while Capon's specification cites literature sources of such information. Eshhar's specification shows the production of chimeric genes encoding scFvR using primers, as listed in Eshhar's Table I. Capon stated that natural genes are isolated and joined using conventional methods, such as the polymerase chain reaction or cloning by primer repair. Capon, like Eshhar, discussed various known procedures for identifying, obtaining, and linking DNA segments, accompanied by experimental examples. The Board did not dispute that persons in this field of science could determine the structure or formula of the linked DNA from the known structure or formula of the components.

The Board stated that "controlling precedent" required inclusion in the specification of the complete nucleotide sequence of "at least one" chimeric gene. Bd. op. at 4. The Board also objected that the claims were broader than

[*18] the specific examples. Eshhar and Capon each responds by pointing to the scientific completeness and depth of their descriptive texts, as well as to their illustrative examples. The Board did not relate any of the claims, broad or narrow, to the examples, but invalidated all of the claims without analysis of their scope and the relation of claim scope to the details of the specifications.

Eshhar and Capon both argue that they have set forth an invention whose scope is fully and fairly described, for the nucleotide sequences of the DNA in chimeric combination is readily understood to contain the nucleotide sequences of the DNA components. Eshhar points to the general and specific description in his specification of known immune-related DNA segments, including the examples of their linking. Capon points similarly to his description of selecting DNA segments that are known to express immune-related proteins, and stresses the existing knowledge of these segments and their nucleotide sequences, as well as the known procedures for selecting and combining DNA segments, as cited in the specification.

Both parties argue that the Board misconstrued precedent, and that precedent does not establish

[*19] a *per se* rule requiring nucleotide-by-nucleotide re-analysis when the structure of the component DNA segments is already known, or readily determined by known procedures.

The Statutory Requirement

The required content of the patent specification is set forth in *Section 112 of Title 35*:

> *§ 112* P1. The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same, and shall set forth the best mode contemplated by the inventor of carrying out his invention.

The "written description" requirement implements the principle that a patent must describe the technology that is sought to be patented; the requirement serves both to satisfy the inventor's obligation to disclose the technologic knowledge upon which the patent is based, and to demonstrate that the patentee was in possession of the invention that is claimed. See *Enzo Biochem, 296 F.3d at 1330* (the written description requirement "is the quid pro quo

[*20] of the patent system; the public must receive meaningful disclosure in exchange for being excluded from practicing the invention for a limited period of time"); *Reiffin v. Microsoft Corp., 214 F.3d 1342, 1345-46 (Fed. Cir. 2000)* (the purpose of the written description requirement "is to ensure that the scope of the right to exclude ...does not overreach the scope of the inventor's contribution to the field of art as described in the patent specification"); *In re Barker, 559 F.2d 588, 592 n. 4 (CCPA 1977)* (the goal of the written description requirement is "to clearly convey the information that an applicant has invented the subject matter which is claimed"). The written description requirement thus satisfies the policy premises of the law, whereby the inventor's technical/scientific advance is added to the body of knowledge, as consideration for the grant of patent exclusivity.

The descriptive text needed to meet these requirements varies with the nature and scope of the invention at issue, and with the scientific and technologic knowledge already in existence. The law must be applied to each invention that enters the patent process, for each patented

[*21] advance is novel in relation to the state of the science. Since the law is applied to each invention in view of the state of relevant knowledge, its application will vary with differences in the state of knowledge in the field and differences in the predictability of the science.

For the chimeric genes of the Capon and Eshhar inventions, the law must take cognizance of the scientific facts. The Board erred in refusing to consider the state of the scientific knowledge, as explained by both parties, and in declining to consider the separate scope of each of the claims. None of the cases to which the Board attributes the requirement of total DNA re-analysis, i.e., Regents v. Lilly, Fiers v. Revel, Amgen, or Enzo Biochem, require a re-description of what was already known. In *Lilly, 119 F.3d at 1567*, the cDNA for human insulin had never been characterized. Similarly in *Fiers, 984 F.2d at 1171*, much of the DNA sought to be claimed was of unknown structure, whereby this court viewed the breadth of the claims as embracing a "wish" or research "plan." In *Amgen, 927 F.2d at 1206*, the court explained that a novel gene was

[*22] not adequately characterized by its biological function alone because such a description would represent a mere "wish to know the identity" of the novel material. In *Enzo Biochem, 296 F.3d at 1326*, this court reaffirmed that deposit of a physical sample may replace words when description is beyond present scientific capability. In *Amgen Inc. v. Hoechst Marion Roussel, Inc., 314 F.3d 1313, 1332 (Fed. Cir. 2003)* the court explained further that the written description requirement may be satisfied "if in the knowledge of the art the disclosed function is sufficiently correlated to a particular, known structure." These evolving principles were applied in *Noelle v. Lederman, 355 F.3d 1343, 1349 (Fed. Cir. 2004)*, where the court affirmed that the human antibody there at issue was not adequately described by the structure and function of the mouse antigen; and in *University of Rochester v. G.D. Searle & Co., 358 F.3d 916, 925-26 (Fed. Cir. 2004)*, where the court affirmed that the description of the COX-2 enzyme did not serve to describe unknown compounds capable of selectively inhibiting the enzyme.

The "written description"

[*23] requirement must be applied in the context of the particular invention and the state of the knowledge. The Board's rule that the nucleotide sequences of the chimeric genes must be fully presented, although the nucleotide sequences of the component DNA are known, is an inappropriate generalization. When the prior art includes the nucleotide information, precedent does not set a *per se* rule that the information must be determined afresh. Both parties state that a person experienced in the field of this invention would know that these known DNA segments would retain their DNA sequences when linked by known methods. Both parties explain that their invention is not in discovering which DNA segments are related to the immune response, for that is in the prior art, but in the novel combination of the DNA segments to achieve a novel result.

The "written description" requirement states that the patentee must describe the invention; it does not state that every invention must be described in the same way. As each field evolves, the balance also evolves between what is known and what is added by each inventive contribution. Both Eshhar and Capon explain that this invention does not concern

[*24] the discovery of gene function or structure, as in Lilly. The chimeric genes here at issue are prepared from known DNA sequences of known function. The Board's requirement that these sequences must be analyzed and reported in the specification does not add descriptive substance. The Board erred in holding that the specifications do not meet the written description requirement because they do not reiterate the structure or formula or chemical name for the nucleotide sequences of the claimed chimeric genes.

Claim Scope

There remains the question of whether the specifications adequately support the breadth of all of the claims that are presented. The Director argues that it cannot be known whether all of the permutations and combinations covered by the claims will be effective for the intended purpose, and that the claims are too broad because they may include inoperative species. The inventors say that they have provided an adequate description and exemplification of their invention as would be understood by persons in the field of the invention. They state that biological properties typically vary, and that their specifications provide for evaluation of the effectiveness

[*25] of their chimeric combinations.

It is well recognized that in the "unpredictable" fields of science, it is appropriate to recognize the variability in the science in determining the scope of the coverage to which the inventor is entitled. Such a decision usually focuses on the exemplification in the specification. See, e.g., *Enzo Biochem, 296 F.3d at 1327-28* (remanding for district court to determine "whether the disclosure provided by the three deposits in this case, coupled with the skill of the art, describes the genera of claims 1-3 and 5"); *Lilly, 119 F.3d at 1569* (genus not described where "a representative number of cDNAs, defined by nucleotide sequence, falling within the scope of the genus" had not been provided); *In re Gosteli, 872 F.2d 1008, 1012 (Fed. Cir. 1989)* (two chemical compounds were insufficient description of subgenus); *In re Smith, 59 C.C.P.A. 1025, 458 F.2d 1389, 1394-95 (CCPA 1972)* (disclosure of genus and one species was not sufficient description of intermediate subgenus); *In re Grimme, 47 C.C.P.A. 785, 274 F.2d 949, 952, 1960 Dec. Comm'r Pat. 123 (CCPA 1960)* (disclosure of single example and

[*26] statement of scope sufficient disclosure of subgenus).

Precedent illustrates that the determination of what is needed to support generic claims to biological subject matter depends on a variety of factors, such as the existing knowledge in the particular field, the extent and content of the prior art, the maturity of the science or technology, the predictability of the aspect at issue, and other considerations appropriate to the subject matter. See, e.g., *In re Wallach, 378 F.3d 1330, 1333-34 (Fed. Cir. 2004)* (an amino acid sequence supports "the entire genus of DNA sequences" that can encode the amino acid sequence because "the state of the art has developed" such that it is a routine matter to convert one to the other); *University of Rochester, 358 F.3d at 925* (considering whether the patent disclosed the compounds necessary to practice the claimed method, given the state of technology); *Singh v. Brake, 317 F.3d 1334, 1343, 48 Fed. Appx. 766 (Fed. Cir. 2002)* (affirming adequacy of disclosure by distinguishing precedent in which the selection of a particular species within the claimed genus had involved "highly unpredictable results").

It

[*27] is not necessary that every permutation within a generally operable invention be effective in order for an inventor to obtain a generic claim, provided that the effect is sufficiently demonstrated to characterize a generic invention. See *In re Angstadt, 537 F.2d 498, 504 (CCPA 1976)* ("The examples, both operative and inoperative, are the best guidance this art permits, as far as we can conclude from the record"). While the Board is correct that a generic invention requires adequate support, the sufficiency of the support must be determined in the particular case. Both Eshhar and Capon present not only general teachings of how to select and recombine the DNA, but also specific examples of the production of specified chimeric genes. For example, Eshhar points out that in Example 1 of his specification the FcR . chain was used, which chain was amplified from a human cDNA clone, using the procedure of Kuster, H. et al., J. Biol. Chem., 265:6448-6451 (1990), which is cited in the specification and reports the complete sequence of the FcRy chain. Eshhar's Example 1 also explains the source of the genes that provide the heavy and light chains of the single chain antibody,

[*28] citing the PhD thesis of Gideon Gross, a co-inventor, which cites a reference providing the complete sequence of the Sp6 light chain gene used to construct the single-chain antibody. Eshhar states that the structure of the Sp6 heavy chain antibody was well known to those of skill in the art and readily accessible on the internet in a database as entry EMBL: MMSP6718. Example 5 at page 54 of the Eshhar specification cites Ravetch et al., J. Exp. Med., 170:481-497 (1989) for the method of producing the CD16 a DNA clone that was PCR amplified; this reference published the complete DNA sequence of the CD16 a chain, as discussed in paragraph 43 of the Eshhar Declaration. Example 3 of the Eshhar specification uses the DNA of the monoclonal anti-HER2 antibody and states that the N29 hybridoma that produces this antibody was deposited with the Collection Nationale de Cultures de Microorganismes, Institut Pasteur, Paris, on August 19, 1992, under Deposit No. CNCM I-1262. It is incorrect to criticize the methods, examples, and referenced prior art of the Eshhar specification as but "a few PCR primers and probes," as does the Director's brief.

Capon's Example 3 provides a detailed description

[*29] of the creation and expression of single chain antibody fused with T-cell receptor zeta chain, referring to published vectors and procedures. Capon, like Eshhar, describes gene segments and their ligation to form chimeric genes. Although Capon includes fewer specific examples in his specification than does Eshhar, both parties used standard systems of description and identification, as well as known procedures for selecting, isolating, and linking known DNA segments. Indeed, the Board's repeated observation that the full scope of all of the claims appears to be "enabled" cannot be reconciled with the Board's objection that only a "general plan" to combine unidentified DNA is presented. See *In re Wands, 858 F.2d 731, 736-37 (Fed. Cir. 1988)* (experimentation to practice invention must not be "undue" for invention to be considered enabled).

The PTO points out that for biochemical processes relating to gene modification, protein expression, and immune response, success is not assured. However, generic inventions are not thereby invalid. Precedent distinguishes among generic inventions that are adequately supported, those that are merely a "wish" or "plan," the words of

[*30] *Fiers v. Revel, 984 F.2d at 1171*, and those in between, as illustrated by *Noelle v. Lederman, 355 F.3d at 1350*; the facts of the specific case must be evaluated. The Board did not discuss the generic concept that both Capon and Eshhar described — the concept of selecting and combining a gene sequence encoding the variable domain of an antibody and a sequence encoding a lymphocyte activation protein, into a single DNA sequence which, upon expression, allows for immune responses that do not occur in nature. The record does not show this concept to be in the prior art, and includes experimental verification as well as potential variability in the concept.

Whether the inventors demonstrated sufficient generality to support the scope of some or all of their claims, must be determined claim by claim. The Board did not discuss the evidence with respect to the generality of the invention and the significance of the specific examples, instead simply rejecting all the claims for lack of a complete chimeric DNA sequence. As we have discussed, that reasoning is inapt for this case. The Board's position that the patents at issue were merely an "invitation to

[*31] experiment" did not distinguish among the parties' broad and narrow claims, and further concerns enablement more than written description. See *Adang v. Fischhoff, 286 F.3d 1346, 1355 (Fed. Cir. 2002)* (enablement involves assessment of whether one of skill in the art could make and use the invention without undue experimentation); *In re Wright, 999 F.2d 1557, 1561 (Fed. Cir. 1993)* (same). Although the legal criteria of enablement and written description are related and are often met by the same disclosure, they serve discrete legal requirements.

The predictability or unpredictability of the science is relevant to deciding how much experimental support is required to adequately describe the scope of an invention. Our predecessor court summarized in *In re Storrs, 44 C.C.P.A. 981, 245 F.2d 474, 478, 1957 Dec. Comm'r Pat. 361 (CCPA 1957)* that "it must be borne in mind that, while it is necessary that an applicant for a patent give to the public a complete and adequate disclosure in return for the patent grant, the certainty required of the disclosure is not greater than that which is reasonable, having due regard to the subject matter involved." This aspect may

[*32] warrant exploration on remand.

In summary, the Board erred in ruling that *§ 112* imposes a *per se* rule requiring recitation in the specification of the nucleotide sequence of claimed DNA, when that sequence is already known in the field. However, the Board did not explore the support for each of the claims of both parties, in view of the specific examples and general teachings in the specifications and the known science, with application of precedent guiding review of the scope of claims.

We remand for appropriate further proceedings.

VACATED AND REMANDED

The Genome Sequence of *Drosophila melanogaster*

Mark D. Adams,[1*] Susan E. Celniker,[2] Robert A. Holt,[1] Cheryl A. Evans,[1] Jeannine D. Gocayne,[1] Peter G. Amanatides,[1] Steven E. Scherer,[3] Peter W. Li,[1] Roger A. Hoskins,[2] Richard F. Galle,[2] Reed A. George,[2] Suzanna E. Lewis,[4] Stephen Richards,[2] Michael Ashburner,[5] Scott N. Henderson,[1] Granger G. Sutton,[1] Jennifer R. Wortman,[1] Mark D. Yandell,[1] Qing Zhang,[1] Lin X. Chen,[1] Rhonda C. Brandon,[1] Yu-Hui C. Rogers,[1] Robert G. Blazej,[2] Mark Champe,[2] Barret D. Pfeiffer,[2] Kenneth H. Wan,[2] Clare Doyle,[2] Evan G. Baxter,[2] Gregg Helt,[6] Catherine R. Nelson,[4] George L. Gabor Miklos,[7] Josep F. Abril,[8] Anna Agbayani,[2] Hui-Jin An,[1] Cynthia Andrews-Pfannkoch,[1] Danita Baldwin,[1] Richard M. Ballew,[1] Anand Basu,[1] James Baxendale,[1] Leyla Bayraktaroglu,[9] Ellen M. Beasley,[1] Karen Y. Beeson,[1] P. V. Benos,[10] Benjamin P. Berman,[2] Deepali Bhandari,[1] Slava Bolshakov,[11] Dana Borkova,[12] Michael R. Botchan,[13] John Bouck,[3] Peter Brokstein,[4] Phillipe Brottier,[14] Kenneth C. Burtis,[15] Dana A. Busam,[1] Heather Butler,[16] Edouard Cadieu,[17] Angela Center,[1] Ishwar Chandra,[1] J. Michael Cherry,[18] Simon Cawley,[19] Carl Dahlke,[1] Lionel B. Davenport,[1] Peter Davies,[1] Beatriz de Pablos,[20] Arthur Delcher,[1] Zuoming Deng,[1] Anne Deslattes Mays,[1] Ian Dew,[1] Suzanne M. Dietz,[1] Kristina Dodson,[1] Lisa E. Doup,[1] Michael Downes,[21] Shannon Dugan-Rocha,[3] Boris C. Dunkov,[22] Patrick Dunn,[1] Kenneth J. Durbin,[3] Carlos C. Evangelista,[1] Concepcion Ferraz,[23] Steven Ferriera,[1] Wolfgang Fleischmann,[5] Carl Fosler,[1] Andrei E. Gabrielian,[1] Neha S. Garg,[1] William M. Gelbart,[9] Ken Glasser,[1] Anna Glodek,[1] Fangcheng Gong,[1] J. Harley Gorrell,[3] Zhiping Gu,[1] Ping Guan,[1] Michael Harris,[1] Nomi L. Harris,[2] Damon Harvey,[4] Thomas J. Heiman,[1] Judith R. Hernandez,[3] Jarrett Houck,[1] Damon Hostin,[1] Kathryn A. Houston,[2] Timothy J. Howland,[1] Ming-Hui Wei,[1] Chinyere Ibegwam,[1] Mena Jalali,[1] Francis Kalush,[1] Gary H. Karpen,[21] Zhaoxi Ke,[1] James A. Kennison,[24] Karen A. Ketchum,[1] Bruce E. Kimmel,[2] Chinnappa D. Kodira,[1] Cheryl Kraft,[1] Saul Kravitz,[1] David Kulp,[6] Zhongwu Lai,[1] Paul Lasko,[25] Yiding Lei,[1] Alexander A. Levitsky,[1] Jiayin Li,[1] Zhenya Li,[1] Yong Liang,[1] Xiaoying Lin,[26] Xiangjun Liu,[1] Bettina Mattei,[1] Tina C. McIntosh,[1] Michael P. McLeod,[3] Duncan McPherson,[1] Gennady Merkulov,[1] Natalia V. Milshina,[1] Clark Mobarry,[1] Joe Morris,[6] Ali Moshrefi,[2] Stephen M. Mount,[27] Mee Moy,[1] Brian Murphy,[1] Lee Murphy,[28] Donna M. Muzny,[3] David L. Nelson,[3] David R. Nelson,[29] Keith A. Nelson,[1] Katherine Nixon,[2] Deborah R. Nusskern,[1] Joanne M. Pacleb,[2] Michael Palazzolo,[2] Gjange S. Pittman,[1] Sue Pan,[1] John Pollard,[1] Vinita Puri,[1] Martin G. Reese,[4] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[3] Bixiang Christopher Shue,[1] Inga Sidén-Kiamos,[11] Michael Simpson,[1] Marian P. Skupski,[1] Tom Smith,[1] Eugene Spier,[1] Allan C. Spradling,[31] Mark Stapleton,[2] Renee Strong,[1] Eric Sun,[1] Robert Svirskas,[32] Cyndee Tector,[1] Russell Turner,[1] Eli Venter,[1] Aihui H. Wang,[1] Xin Wang,[1] Zhen-Yuan Wang,[1] David A. Wassarman,[33] George M. Weinstock,[3] Jean Weissenbach,[14] Sherita M. Williams,[1] Trevor Woodage,[1] Kim C. Worley,[3] David Wu,[1] Song Yang,[2] Q. Alison Yao,[1] Jane Ye,[1] Ru-Fang Yeh,[19] Jayshree S. Zaveri,[1] Ming Zhan,[1] Guangren Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Xiangqun H. Zheng,[1] Fei N. Zhong,[1] Wenyan Zhong,[1] Xiaojun Zhou,[3] Shiaoping Zhu,[1] Xiaohong Zhu,[1] Hamilton O. Smith,[1] Richard A. Gibbs,[3] Eugene W. Myers,[1] Gerald M. Rubin,[34] J. Craig Venter[1]

The fly *Drosophila melanogaster* is one of the most intensively studied organisms in biology and serves as a model system for the investigation of many developmental and cellular processes common to higher eukaryotes, including humans. We have determined the nucleotide sequence of nearly all of the ~120-megabase euchromatic portion of the *Drosophila* genome using a whole-genome shotgun sequencing strategy supported by extensive clone-based sequence and a high-quality bacterial artificial chromosome physical map. Efforts are under way to close the remaining gaps; however, the sequence is of sufficient accuracy and contiguity to be declared substantially complete and to support an initial analysis of genome structure and preliminary gene annotation and interpretation. The genome encodes ~13,600 genes, somewhat fewer than the smaller *Caenorhabditis elegans* genome, but with comparable functional diversity.

The annotated genome sequence of *Drosophila melanogaster*, together with its associated biology, will provide the foundation for a new era of sophisticated functional studies (*1–3*). Because of its historical importance, large research community, and powerful research tools, as well as its modest genome size, *Drosophila* was chosen as a test system to explore the applicability of whole-genome shotgun (WGS) sequencing for large and complex eukaryotic genomes (*4*). The groundwork for this project was laid over many years by the fly research community, which has molecularly characterized ~2500 genes; this work in turn has been supported by nearly a century of genetics (*5*). Since *Drosophila* was chosen in 1990 as one of the model organisms to be studied under the auspices of the federally funded Human Genome Project, genome projects in the United States, Europe, and Canada have produced a battery of genome-wide resources (Table 1). The Berkeley and European *Drosophila* Genome Projects (BDGP and EDGP) initiated genomic sequencing (Tables 1 to 3) and finished 29 Mb. The bacterial artificial chromosome (BAC) map and other genomic resources available for *Drosophila* serve both as an independent confirmation of the assembly of data from the shotgun strategy and as a set of resources for further biological analysis of the genome.

The *Drosophila* genome is ~180 Mb in size, a third of which is centric heterochromatin (Fig. 1). The 120 Mb of euchromatin is on two large autosomes and the X chromosome; the small fourth chromosome contains only ~1 Mb of euchromatin. The heterochromatin consists mainly of short, simple sequence elements repeated for many megabases, occasionally interrupted by inserted transposable elements, and tandem arrays of ribosomal RNA genes. It is known that there are small islands of unique sequence embedded within heterochromatin—for example, the mitogen-activated protein kinase gene *rolled* on chromosome 2, which is flanked on each side by at least 3 Mb of heterochromatin. Unlike the *C. elegans* genome, which can be completely cloned in yeast artificial chromosomes (YACs), the simple sequence repeats are not stable in YACs (*6*) or other large-insert cloning systems. This has led to a functional definition of the euchromatic genome as that portion of the genome that can be cloned stably in BACs. The euchromatic portion of the genome is the subject of both the federally funded *Drosophila* sequencing project and the work presented here. We began WGS sequencing of *Drosophila* less than 1 year ago, with two major goals: (i) to test the strategy on a large and complex eukaryotic genome as a prelude to sequencing the human genome, and (ii) to provide a complete, high-quality genomic sequence to the *Drosophila* research community so as to advance research in this important model organism.

WGS sequencing is an effective and efficient way to sequence the genomes of prokaryotes, which are generally between 0.5 and 6 Mb in size (*7*). In this strategy, all the DNA of an organism is sheared into segments a few thousand base pairs (bp) in length and cloned directly into a plasmid vector suitable for DNA sequencing. Sufficient DNA sequencing is performed so that each base pair is covered numerous times, in fragments of ~500 bp. After sequencing, the fragments are assembled in overlapping segments to reconstruct the complete genome sequence.

In addition to their much larger size, eukaryotic genomes often contain substantial amounts of repetitive sequence that have the potential to interfere with correct sequence assembly. Weber and Myers (*8*) presented a theoretical analysis of WGS sequencing in which they examined the impact of repetitive sequences, discussed experimental strategies to mitigate their effect on sequence assembly, and suggested that the WGS method could be applied effectively to large eukaryotic genomes. A key component of the strategy is obtaining sequence data from each end of the cloned DNA inserts; the juxtaposition of these end-sequences ("mate pairs") is a critical element in producing a correct assembly.

Genomic Structure

WGS libraries were prepared with three different insert sizes of cloned DNA: 2 kb, 10 kb, and 130 kb. The 10-kb clones are large enough to span the most common repetitive sequence elements in *Drosophila*, the retrotransposons. End-sequence from the BACs provided long-range linking information that was used to confirm the overall structure of the assembly (*9*). More than 3 million sequence reads were obtained from whole-genome libraries (Fig. 2 and Table 2). Only ~2% of the sequence reads contained heterochromatic simple sequence repeats, indicating that the heterochromatic DNA is not stably cloned in the small-insert vectors used for the WGS libraries. A BAC-based physical map spanning >95% of the euchromatic portion of the genome was constructed by screening a BAC library with sequence-tagged site (STS) markers (*10*). More than 29 Mb of high-quality finished sequence has been completed from BAC, P1, and cosmid clones, and draft sequence data (~1.5× average coverage) were obtained from an additional 825 BAC and P1 clones spanning in total >90% of the genome (Table 3). The clone-based draft sequence served two purposes: It improved the likelihood of accurate assembly, and it allowed the identification of templates and primers for filling gaps that remain after assembly. An initial assembly was performed using the WGS data and BAC end-sequence [WGS-only assembly (*4*)]; subsequent assemblies included the clone-based draft sequence data (joint assembly). Figure 3 and Table 3 illustrate the status of the euchromatic sequence resulting from each of these assemblies and the current status following the directed gap closure completed to date. The sequence assembly process is described in detail in an accompanying paper (*11*).

Assembly resulted in a set of "scaffolds." Each scaffold is a set of contiguous sequences (contigs), ordered and oriented with respect to one another by mate-pairs such that the gaps between adjacent contigs are of known size and are spanned by clones with end-sequences flanking the gap. Gaps within scaffolds are called sequence gaps; gaps between scaffolds are called "physical gaps" because there are no clones identified spanning the gap. Two methods were used to map the scaffolds to chromosomes: (i) cross-referencing between STS markers present in the assembled sequence and the BAC-based STS content map, and (ii) cross-referencing between assembled sequence and shotgun sequence data obtained from individual tiling-path clones selected from the BAC physical map. The mapped scaffolds from the joint assembly, totaling 116.2 Mb after initial

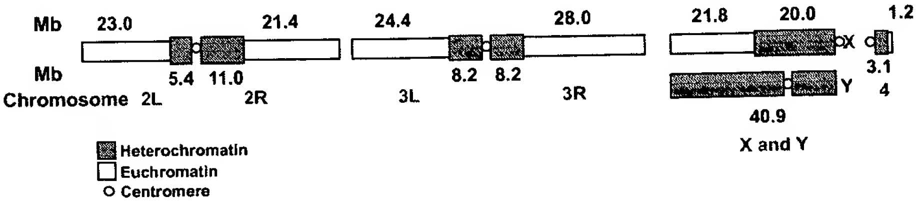


Fig. 1. Mitotic chromosomes of *D. melanogaster*, showing euchromatic regions, heterochromatic regions, and centromeres. Arms of the autosomes are designated 2L, 2R, 3L, 3R, and 4. The euchromatic length in megabases is derived from the sequence analysis. The heterochromatic lengths are estimated from direct measurements of mitotic chromosome lengths (*67*). The heterochromatic block of the X chromosome is polymorphic among stocks and varies from one-third to one-half of the length of the mitotic chromosome. The Y chromosome is nearly entirely heterochromatic.

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]Berkeley *Drosophila* Genome Project (BDGP), Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [4]BDGP, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [5]European Molecular Biology Laboratory (EMBL)–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [6]Neomorphic Inc., 2612 Eighth Street, Berkeley, CA 94710, USA. [7]GenetixXpress Pty. Ltd., 78 Pacific Road, Palm Beach, Sydney, NSW 2108, Australia. [8]Department of Medical Informatics, IMIM–UPF C/Dr. Aiguader 80, 08003 Barcelona, Spain. [9]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA. [10]Department of Genetics, Box 8232, Washington University Medical School, 4566 Scott Avenue, St. Louis, MO 63110, USA. [11]Institute of Molecular Biology and Biotechnology, Forth, Heraklion, Greece. [12]European *Drosophila* Genome Project (EDGP), EMBL, Heidelberg, Germany. [13]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94710, USA. [14]Genoscope, 2 rue Gaston Crémieux, 91000 Evry, France. [15]Section of Molecular and Cellular Biology, University of California, Davis, CA 95618, USA. [16]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK. [17]EDGP, Rennes University Medical School, UPR 41 CNRS Recombinaisons Genetiques, Faculte de Medicine, 2 av. du Pr. Leon Bernard, 35043 Rennes Cedex, France. [18]Department of Genetics, Stanford University, Palo Alto, CA 94305, USA. [19]Department of Statistics, University of California, Berkeley, CA 94720, USA. [20]EDGP, Centro de Biologia Molecular Severo Ochoa, CSIC, Universidad Autónoma de Madrid, 28049 Madrid, Spain. [21]MBVL, Salk Institute, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA. [22]Department of Biochemistry and Center for Insect Science, University of Arizona, Tucson, AZ 85721, USA. [23]EDGP, Montpellier University Medical School, Institut de Genetique Humaine, CNRS (CRBM), 114 rue de la Cardonille, 34396 Montpellier Cedex 5, France. [24]Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health (NIH), Bethesda, MD 20892, USA. [25]Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montreal, Quebec, Canada. [26]The Institute for Genomic Research, Rockville, MD 20850, USA. [27]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA. [28]EDGP, Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [29]Department of Biochemistry, University of Tennessee, Memphis, TN 38163, USA. [30]EDGP, Department of Anatomy and Physiology, University of Dundee, Dundee DD1 4HN, UK, and Department of Biological Sciences, Open University, Milton Keynes MK7 6AA, UK. [31]HHMI/Embryology, Carnegie Institution of Washington, Baltimore, MD 21210, USA. [32]Motorola BioChip Systems, Tempe, AZ 85284, USA. [33]Cell Biology and Metabolism Branch, National Institute of Child Health and Human Development, NIH, Bethesda, MD 20892, USA. [34]Howard Hughes Medical Institute, BDGP, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed.

gap closure, were deposited in GenBank (accession numbers AE002566–AE003403) and form the basis for the analysis described in this article.

The WGS-only assembly resulted in 50 scaffolds spanning 114.8 Mb that could be placed unambiguously onto chromosomes solely on the basis of their STS content (labeled "D" in Fig. 3). The joint assembly included clone-based sequence, but no specific advantage was taken of the location information of each clone-based read by the whole-genome assembly algorithm. Nonetheless, the clone-based sequence from BACs in the physical map allowed placement of an additional 84 small scaffolds (1.4 Mb) on chromosome arms in the joint assembly (labeled "C" in Fig. 3). As shown in Fig. 3, a few large scaffolds in each assembly span a large portion of each chromosome arm, with a number of additional smaller scaffolds located at the centromeric end, except on the right arm of chromosome 3. Nearly all of the scaffolds added to chromosomes in the joint assembly, relative to the WGS-only assembly, are adjacent to the centric heterochromatin, which demonstrates the utility of the physical map in these regions. The density of transposable elements (labeled "A" in Fig. 3) increases markedly in the transition zone between euchromatin and heterochromatin, as discussed below. An additional 704 scaffolds in the joint assembly, equivalent to 3.8 Mb, could not be placed with accuracy on the genome. Most of these do not match clone-based sequence from the physical map, and therefore they most likely represent small islands of unique sequence embedded within regions of heterochromatin. Because of the instability of the surrounding genomic regions, these sequences would not have been obtained through a sequencing approach that was dependent on cloning in large-insert vectors.

Among the 134 mapped scaffolds, there were 1636 contigs after assembly (hence 1630 gaps, considering that there are six linear chromosome arm segments to be assembled). On the major autosomes, there are five physical gaps in the BAC map, three of which are near a centromere or telomere (*10*). Because the WGS approach did not span these gaps, they likely contain unclonable regions. Most gaps on the autosomes—including gaps between scaffolds—were therefore cloned in either WGS clones or BAC subclones used for clone-based draft sequencing and are considered sequence gaps. Directed gap closure was done through use of several resources, including whole BAC clones, plasmid subclones, and M13 subclones from the Lawrence Berkeley National Laboratory (LBNL) and Baylor College of Medicine centers' draft sequence of BAC and P1 clones; 10-kb subclones from the whole-genome libraries; and polymerase chain reaction (PCR) from genomic DNA (*12*). The average size of the gaps filled to date is 771 bp (their predicted size was 757 bp); the predicted size of the remaining gaps is 2120 bp. Table 3 provides details of the status of each chromosome arm as of 3 March 2000.

The accuracy of the assembly was measured in several ways, as described (*11*). In summary, the scaffold sequences agree very well with the BAC-based STS content map and with high-quality finished sequence. In the 7 Mb of the genome where very high-quality sequence was

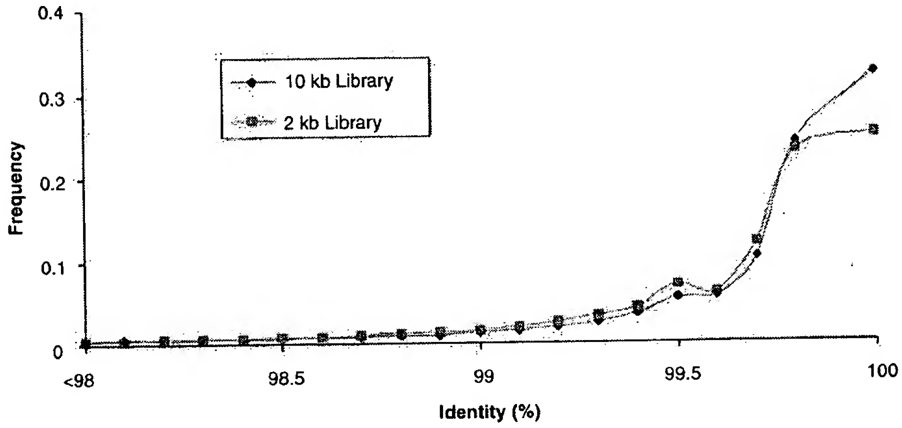


Fig. 2. Accuracy of sequence reads from ABI Prism 3700 DNA analyzer. A database of BAC and P1 clone sequences from BDGP finished to high accuracy ($P_{sum} > 100,000$, indicating less than one error predicted per 100,000 bases) was constructed. Trimmed WGS sequence reads matching these BAC and P1 clones were identified by BLAST. The first high-scoring pair (HSP) with a full-length match was used. Identity is the percentage of matched nucleotides in the alignment; 49,756 sequence reads from 2-kb libraries and 23,455 reads from 10-kb libraries matched these BAC and P1 sequences. The average trimmed read length of sequences from 2-kb and 10-kb clones was 570 bp and 567 bp, respectively.

Table 1. Genomic resources for *Drosophila*.

Type	Description	Resolution	Contribution	Source and reference
BAC-based STS content map	STS content map constructed by screening ~23× genome coverage of BAC clones; a tiling path of BACs spanning each chromosome arm was selected	50 kb	Location of whole-genome scaffolds to chromosomes; confirmation of accuracy of assembly	BDGP [chromosomes 2 and 3 (*10*)], EDGP [X chromosome (*69*), www.dundee.ac.uk/anatphys/robert/Xdivs/MapIntro.htm], University of Alberta [chromosome 4 (*70*)]
Polytene map	Tiling-path BACs hybridized to polytene chromosomes	30 kb	Location of STSs and BACs to chromosomes; validation of BAC map	See (*10*)
BAC end-sequence	~500 bp of sequence from each end of a BAC clone	Two reads per ~130 kb	Long-range association of sequence contigs	Genoscope (www.genoscope.fr)
Finished clone-based sequence	BAC, P1, and cosmid clones completely sequenced to high accuracy	~29 Mb of total sequence	Assessment of accuracy of Celera sequence and assembly	LBNL (26 Mb), EDGP [3 Mb (*69*)]
Draft sequence from mapped BACs	≥1.5× shotgun sequence coverage of 825 clones from the tiling path of BAC and P1 clones	384 reads distributed across ~160 kb	Location of sequence contigs to a small genomic region; templates for gap closure	LBNL, Baylor College of Medicine

available for comparison, the accuracy of the assembled sequence was 99.99% in nonrepetitive regions. In the ~2.5% of the region comprising the most highly repetitive sequences, the accuracy was 99.5%.

Heterochromatin-euchromatin transition zone. The genomes of eukaryotes generally contain heterochromatic regions surrounding the centromeres that are intractable to all current sequencing methods. In *Drosophila*, ~60 Mb of the 180-Mb genome consists of centric heterochromatin, which is composed primarily of simple sequence satellites, transposons, and two large blocks of ribosomal RNA genes (*13*). We examined the sequence organization at boundaries between euchromatin and centric heterochromatin in two regions, one in division 20 on the X chromosome and the other in division 40 on the left arm of chromosome 2. On the X chromosome, gene density in division 20 drops abruptly—to two genes in 400 kb around *folded gastrulation*—and then rises to 11 genes in 130 kb. Next, at least 10 Mb of largely satellite DNA sequences and the ribosomal RNA gene cluster are located just distal to the centromere itself. On the left arm of chromosome 2, a similar situation exists: There is a normal gene density in division 39, followed by only two genes in 350 kb near *teashirt* in division 40, then by a 200-kb region containing 10 genes. These transition zones between euchromatin and heterochromatin contain many previously unknown genes, including counterparts to human cyclin K and mouse Krox-4. None of the 11 genes proximal to *teashirt* and only one of the 10 genes proximal to *folded gastrulation* was known previously.

What is the nature of the sequence in the gene-poor regions? The most common sequences by far were transposons, consistent with previous small-scale analyses (*14*). These include several new elements similar to transposons in other species, as well as the ~50 transposon classes previously characterized in *Drosophila*. Some short runs of satellite sequences are present, but it has not been determined whether they might have been truncated during cloning. In addition, at least 110 other simple repeat classes were identified, some of which are distributed widely outside of heterochromatin.

Criteria for describing the completion status of a eukaryotic genome. Because of the unclonable repetitive DNA surrounding the centromeres, it is highly unlikely that the genomic sequence of chromosomes from eukaryotes such as *Drosophila* or human will ever be "complete." It is therefore necessary to provide an assessment of the contiguity and accuracy of the sequence. Table 4 lists several objective parameters by which the status can be judged and by which improvements in future releases can be measured. We have termed the version of the sequence associated with this publication "Release 1" and intend to make regular future releases as gaps are filled and overall sequence accuracy is increased.

One measure of the completeness of the assembled sequence is the extent to which previously described genes can be found. An analysis of the 2783 *Drosophila* genes with some sequence information that have been compiled by FlyBase (*15*) resulted in identification of 2778 on the scaffold sequence. All of the remainder are found in unscaffolded sequence. The remaining six were all cloned by degenerate PCR, and it is possible that some or all of these genes are incorrectly ascribed to *Drosophila* (*16*). Of the base pairs represented in the 2778 genes, 97.5% are present in the assembled sequence.

Annotation

The initial annotation of the assembled genome concentrated on two tasks: prediction of transcript and protein sequence, and prediction of function for each predicted protein. Computational approaches can aid each task, but biologists with expertise in particular fields are required for the results to have the most consistency, reliability, and utility. Because the breadth of expertise necessary to annotate a complete genome does not exist in any single individual or organization, we hosted an "Annotation Jamboree" involving more than 40 scientists from around the world, primarily from the *Drosophila* research community. Each was responsible for organizing and interpreting the gene set for a given protein family or biological process. Over a 2-week period, jambo

Table 2. Source of data for assembly: Whole-genome shotgun sequencing. See (*65*) for more information about library construction and sequencing.

Vector	Insert size (kbp)	Paired sequences	Total sequences	Clone coverage	Sequence coverage
High-copy plasmid	2	732,380	1,903,468	11.2×	7.3×
Low-copy plasmid	10	548,974	1,278,386	42.2×	5.4×
BAC	130	9,869	19,738	11.4×	0.07×
Total		1,290,823	3,201,592	64.8×	12.8×

Table 3. BAC and P1 clone-based sequencing. EDGP, European *Drosophila* Genome Project; BCM, Baylor College of Medicine; LBNL, Lawrence Berkeley National Laboratory (BCM and LBNL are the genomic sequencing centers of the BDGP).

Clone-based genomic sequencing								Gap closure: current status		
				Draft sequence in joint assembly [BACs, (P1s)]†				Percentage of DNA sequence in contigs greater than		
Chromosomal region	Group	Size	Finished sequence (Mb)	Clones	Average coverage	Total sequenced BACs (P1s) in joint assembly	Additional sequenced BACs in tiling path	30 kb	100 kb	1 Mb
X (1–3)	EDGP	3	2.5	0		0	0	79.4	32.7	0
X (4–11)	BCM	8.8	0.1*	0		1	72			
X (12–20)	LBNL	10	0	71	2.3×	71	10			
2L	LBNL	23	14.0	103 (8)	1.6× (5.3×)	119 (202)	2	97.8	91.4	16.9
2R	LBNL	21.4	8.8	159 (32)	1.3× (4.7×)	157 (186)	0	96.4	90.6	32.8
3L	BCM	24.4	0.1	166	1.3×	170	50	95.1	77.7	0
3L	LBNL		2.1	22 (7)	1.7× (2.5×)	20 (32)	0			
3R	LBNL	28	2.1	259 (9)	1.2× (2×)	264 (27)	0	98.5	92.6	3.6
4	LBNL	1.2	0	16	1.4×	15	1	85.6	43.5	0
Total		120	29.7	796 (56)		817 (447)	135	93.7	77.5	9.9

*Sequenced at LBNL. †A tiling path of clones spanning 97% of the euchromatic portion of the genome was selected from the genome physical maps (*10*) for clone-based sequencing. The data include sequence that has been generated since the beginning of the publicly funded (BDGP and EDGP) genome sequencing projects. Tiling path clone identities were verified by screening the shotgun sequence for expected STS and BAC end-sequences, sequenced genes with known map locations from genes (and regions flanking P insertions), and sequences of neighboring tiling path clones. The average size of BAC clones in the tiling path is 163 kb. Sequencing methods are described in (*66*).

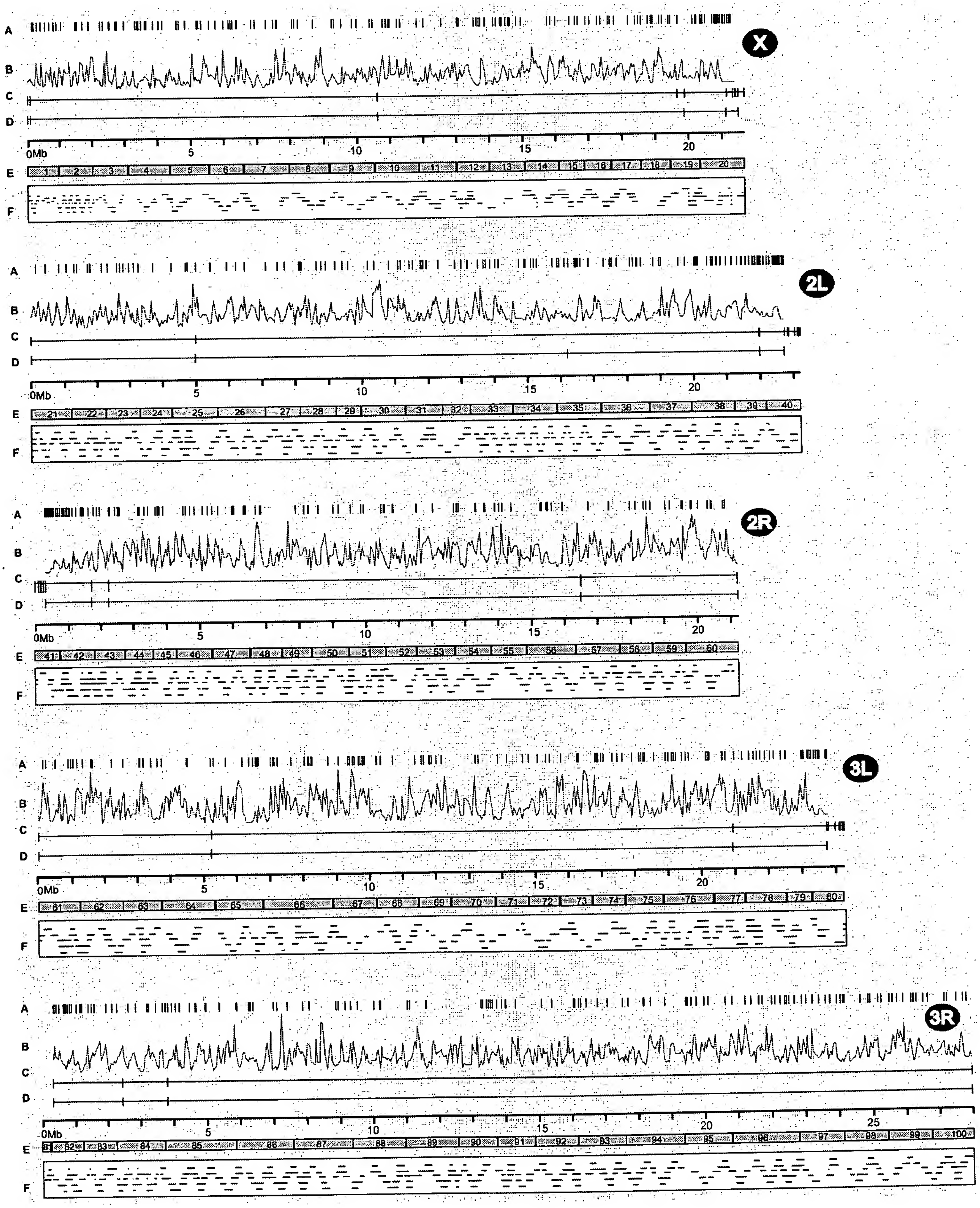


Fig. 3. Assembly status of the *Drosophila* genome. Each chromosome arm is depicted with information on content and assembly status: **(A)** transposable elements, **(B)** gene density, **(C)** scaffolds from the joint assembly, **(D)** scaffolds from the WGS-only assembly, **(E)** polytene chromosome divisions, and **(F)** clone-based tiling path. Gene density is plotted in 50-kb windows; the scale is from 0 to 30 genes per 50 kb. Gaps between scaffolds are represented by vertical bars in (C) and (D). Clones colored red in the tiling path have been completely sequenced; clones colored blue have been draft-sequenced. Gaps shown in the tiling path do not necessarily mean that a clone does not exist at that position, only that it has not been sequenced. Each chromosome arm is oriented left to right, such that the centromere is located at the right side of X, 2L, and 3L and the left side of 2R and 3R.

ree participants worked to define genes, to classify them according to predicted function, and to begin synthesizing information from a genome-wide perspective.

For definition of gene structure, we relied on the use of different gene-finding approaches: the gene-finding programs Genscan (*17*) and a version of Genie that uses expressed sequence tag (EST) data (*18*), plus the results of complementary DNA (cDNA) and protein database searches, followed by review by human annotators (*19*). Genscan predicted 17,464 genes, and Genie predicted 13,189. We believe that the lower estimate is more accurate, because in a test that used the extensively studied and annotated 2.9-Mb *Adh* region (*3*), the Genie predictions were closer to the number of experimentally determined genes; Genscan predicted far too many (*20*). This is likely because Genie was optimized for *Drosophila*, whereas Genscan parameters suitable for *Drosophila* gene-finding are not available.

Results of the computational analyses were presented to annotators by means of a custom visualization tool that allowed annotators to define transcripts on the basis of EST (*21*) and protein sequence similarity information, Genie predictions, and Genscan predictions, in decreasing order of confidence. The present annotation of the *Drosophila* genome predicts 13,601 genes, encoding 14,113 transcripts through alternative splicing in some genes. The number of alternative splice forms that can be annotated is limited by the available cDNA data and is a substantial underestimate of the total number of alternatively spliced genes. More than 10,000 genes with database matches were reviewed manually. The remaining ~3000 genes were predicted by Genie but have no database matches that can be used to refine intron-exon boundaries. Genes predicted by Genscan that did not overlap Genie predictions or database matches were not included in the set of predicted proteins. Table 5 summarizes the evidence for these genes: 38% of the Genie predictions are supported by evidence from both EST and protein matches, 27% by ESTs alone, and 12% by protein matches alone. Altogether there are EST matches for 65% of the genes, but nearly half of the total ESTs match only 5% of the genes; 23% of the predicted proteins do not match sequences from other organisms or *Drosophila* ESTs. This set of annotations is considered provisional and will improve as additional full-length cDNA sequence and functional information becomes available for each gene. Figure 4 provides a graphical overview of the gene content of the fly.

Genes were classified according to a functional classification scheme called Gene Ontology (GO). The GO project (*22*) is a collaboration among FlyBase, the *Saccharomyces* Genome Database (*23*), and Mouse Genome Informatics (*24*). It consists of a set of controlled vocabularies providing a consistent description of gene products in terms of their molecular function, biological role, and cellular location. At the time of our annotation, proteins encoded by 1539 *Drosophila* genes had already been annotated by FlyBase using ~1200 different GO classifications. In addition, a set of 718 proteins from *S. cerevisiae* and 1724 proteins from mouse had been annotated and placed into GO categories. Predicted *Drosophila* genes and gene products were used as queries against a database made up of the sequences of these three sets of proteins (by BLASTX or BLASTP) (*25*) and grouped on the basis of the GO classification of the proteins matched. About 7400 transcripts have been assigned to 39 major functional categories, and about 4500 have been assigned to 47 major process categories (Table 6).

The largest predicted protein is Kakapo, a cytoskeletal linker protein required for adhesion between and within cell layers, with 5201 amino acids; the smallest is the 21–amino acid ribosomal protein L38. There are 56,673 predicted exons, an average of four per gene, occupying 24.1 Mb of the 120-Mb euchromatic sequence total. The size of the average predicted transcript is 3058 bp. There was a systematic underprediction of 5′ and 3′ untranslated sequence as a result of less than complete EST coverage and the inability of gene-prediction programs to predict the noncoding regions of transcripts, so the number of exons and introns and the average transcription unit size are certain to be underestimates. There are at least 41,000 introns, occupying 20 Mb of sequence. Intron sizes in *Drosophila* are heterogeneous, ranging from 40 bp to more than 70 kb, with a clear peak between 59 and 63 bp (*26*). The average number of exons is four, although this is an underestimate because of a systematic underprediction of 5′ and 3′ untranslated exons. We identified 292 transfer RNA genes and 26 genes for spliceosomal small nuclear RNAs (snRNAs). We did not attempt to predict other noncoding RNAs.

Table 4. Measures of completion. Analyses supporting many of these values are found in (*11*).

Number of scaffolds mapped to chromosome arms	134
Number of scaffolds not mapped to chromosomes	704
Number of base pairs in scaffolds mapped to chromosome arms	116.2 Mb
Number of base pairs in scaffolds not mapped to chromosome arms	3.8 Mb
Largest unmapped scaffold	64 kb
Percentage of total base pairs in mapped scaffolds >100 kb	98.2%
Percentage of total base pairs in mapped scaffolds >1 Mb	95.5%
Percentage of total base pairs in mapped scaffolds >10 Mb	68.0%
Number of gaps remaining among mapped scaffolds	1299
Base pair accuracy against LBNL BACs (nonrepetitive sequence)	99.99%
Known genes accounted for in scaffold set	99.7%

The total number of protein-coding genes, 13,601, is less than that predicted for the worm *C. elegans* (*27*) (18,425; WormPep 18, 11 October 1999) and far less than the ~27,000 estimated for the plant *Arabidopsis thaliana* (*28*). The average gene density in *Drosophila* is one gene per 9 kb. There is substantial variation in gene density, ranging from 0 to nearly 30 genes per 50 kb, but the gene-rich regions are not clustered as they are in *C. elegans*. Regions of high gene density correlate with G+C-rich sequences. In the ~1 Mb adjacent to the centric heterochromatin, both G+C content and gene density decrease, although there is not a marked decrease in EST coverage as has been seen in *A. thaliana* (*28*).

Genomic Content

The genomic sequence has shed light on some of the processes common to all cells, such as replication, chromosome segregation, and iron metabolism. There are also new findings about important classes of chromosomal proteins that allow insights into gene regulation and the cell cycle. Overall, the correspondence of *Drosophila* proteins involved in gene expression and metabolism to their human counterparts reaffirms that the fly represents a suitable experimental platform for the examination of human disease networks involved in replication, repair, translation, and the metabolism of drugs and toxins. In an accompanying manuscript (*29*), the protein complement of *Drosophila* is compared to those of the two eukaryotes with complete genome sequences, *C. elegans* and *S. cerevisiae*, and other developmental and cell biological processes are discussed.

Replication. Genes encoding the basic DNA replication machinery are conserved among eukaryotes (*30*); in particular, all of the proteins known to be involved in start site recognition are encoded by single-copy genes in the fly. These include members of the six-subunit heteromeric origin recognition complex (ORC) (*31*), the MCM helicase complex (*32*), and the regulatory factors CDC6 and CDC45, which are thought to determine processing of pre-initiation complexes. The fly ORC3 and ORC6 proteins, for example, share close sequence similarity with vertebrate proteins, but not only are they highly divergent relative to yeast ORCs, they have no obvious counterparts in the worm. It is striking that the ORC genes exist as single copies, given the orthologous functions for some of the subunits in other processes (*33*). It had been considered possible that a large family of ORCs, each with a different binding specificity, might account for different origin usage in development. Clearly, given the single-copy ORC genes, other as-yet-undiscovered cis-acting elements and trans-acting factors participate in developmentally regulated processes such as switches in origin usage, gene amplification, and specialized replication of euchromatin in certain endocycles. In contrast, the fly has two distinct homologs of the proliferating cell nuclear antigen (PCNA), the processivity factor for the DNA polymerases (δ and ε) involved in chain elongation. Human PCNA is blocked from interaction with the replication enzymes by the checkpoint regulator p21 in response to DNA damage (*34*); perhaps one of the fly PCNA proteins is immune to such regulation and is thus left active for repair or replication.

Chromosomal proteins. Analysis of protein families involved in chromosome inheritance reveals both expected findings and some surprises. As expected, the fly has all four members of the conserved SMC family involved in sister chromatid cohesion, condensation, DNA repair, and dosage compensation (*35*). The fly also contains at least one ortholog of each of the MAD/Bub metaphase-anaphase checkpoint proteins that are conserved from yeast to mammals. However, *Drosophila* does not appear to have orthologs to most of the proteins identified previously in mammals or yeast that are associated with centromeric DNA, such as the CENP-C/MIF-2 family and the yeast CBF3 complex (*36*). One exception is the presence of a histone H3-like protein that shares sequence similarity with mammalian CENP-A, a centromere-specific H3-like protein. There are at least nine histone acetyltransferases (HATs) and five histone deacetylases (HDACs), which are involved in regulating chromatin structure (*37*); only three of each have been reported previously. There are also 17 members of the SNF2 adenosine triphosphatase (ATPase) family, which represent 9 of the 10 known subfamilies. Many of these ATPases are involved in chromatin remodeling (*38*). The fly also contains at least 14 proteins with chromodomains (*39*), six of which are new, including two HP1-related proteins. Although many of these chromodomain-containing proteins have orthologs in vertebrates, only one (CHD1) appears in yeast, flies, and vertebrates. There are also at least 13 bromodomain-containing proteins, seven of which are new; the bromodomain may interact with the acetylated NH$_2$-terminus of histones and is involved in chromatin remodeling and gene silencing (*40*). Only three of these appear to have counterparts in yeast. Furthermore, *Drosophila* telomeres lack the simple repeats that are characteristic of most eukaryotic telomeres (*41*), and the known telomerase components of vertebrates, for example, are absent from flies. The fly does, however, contain five proteins that are close relatives of the yeast and human SIR2 telomere silencing proteins.

DNA repair. The importance of DNA repair in maintaining genomic integrity is reflected in the conservation of most proteins implicated in the major defined pathways of eukaryotic DNA repair. However, there are some notable absences. For example, no convincing homologs can be found for the genes encoding the RAD7, RAD16, RAD26 (CSB/ERCC6), and RAD28 (CSA) proteins, which are implicated in strand-specific modes of repair in yeast and/or mammalian systems. In base excision repair processes, 3-methyladenine glycosylase and uracil-DNA-glycosylase are absent, although the latter function is likely fulfilled by the G/T mismatch-specific thymine DNA glycosylase (*42*). In the damage bypass pathway, sequences encoding homologs of DNA polymerase ζ (yeast Rev3p/*Drosophila* mus205) and Rev1p are present, although a REV7 homolog is not found. As in humans and worms, two members of the RAD30 (polymerase η) gene family are present. In the mismatch repair system, only two proteins related to *Escherichia coli* mutS are predicted, rather than the usual family of five or more members. The previously reported Msh2p homolog (*43*) is present, as is a sequence most closely resembling Msh6p. Budding yeast and humans possess additional members of the mutS gene family that are proposed to function in partially redundant pathways of mismatch repair (MSH3) and in meiotic recombination (MSH4 and MSH5), suggesting either that the *Drosophila* mutS homologs have reduced specificity or that alternative proteins are fulfilling these roles in the fly. In the recombinational repair pathway, two additional members of the recA/RAD51 gene family are identified, bringing the total to four. However, no member of the RAD52/RAD59 family is present. One additional member of the recQ/SGS1 helicase family was identified, in addition to the two already noted (*44*); the new protein is most similar to human RecQ4. Finally, with respect to nonhomologous end joining, *Drosophila* joins the list of invertebrate species that lack an apparent DNA-PK catalytic subunit, although both Ku subunits and DNA ligase 4 are present. We conclude that most major components of the repair network in flies have been uncovered. If more are present, either they have diverged so far that they are unrecognizable by BLAST searches, or the systems have become degenerate (that is, other network components are fulfilling the same roles).

Transcription. Gene regulation has traditionally been singled out as one of the primary bases for the generation of evolutionary diversity. How has the core transcriptional machinery changed in different phyla? *Drosophila* core RNA polymerase II and some general transcription factors (TFIIA-H, TFIIIA, and TFIIIB) are similar in composition to those of both mammals and yeast (*45*). In contrast, core RNA polymerases I and III, TBP (TATA-binding protein)–containing complexes for class I, class II, and snRNA genes (TBP-associated factors TAF$_{I}$ and TAF$_{II}$, and SNAP$_{C}$, respectively), TFIIIC, and SRB/mediator vary greatly in composition in *Drosophila* and mammals relative to yeast (*46*). The RNA polymerase I transcription factors of flies and mammals have clear amino acid conservation; yeast RNA polymerase I factors do not appear to be related to them. For example, the mammalian promoter interacting factors UBF and TIF-1A are present in *Drosophila* but not in yeast, and yeast UAF subunits are absent in *Drosophila* and apparently absent in mammals. Furthermore, of the three TAF$_{I}$s in the human selectivity factor 1, the mouse transcriptional initiation factor IB, and the yeast core factor complexes, only the human/mouse TAF$_{I}$63/TAF$_{I}$68 subunit is conserved in the fly. Similarly, *Drosophila* encodes three of the five mammalian SNAP$_{C}$ subunits (SNAP43, 50, and 190) for which no homologs exist in the yeast genome.

In addition to the family of previously described TBPs (*47*), the fly contains multiple forms of several ubiquitous TAF$_{II}$s (TAF$_{II}$30β, TAF$_{II}$60, and TAF$_{II}$80) (*46*). This raises the possibility that a variety of TFIID complexes evolved in metazoan organisms to regulate gene expression patterns associated with development and cellular differentiation. The constellation of factors that interact with RNA polymerase II in *Drosophila* may also contribute to this regulation, because *Drosophila* contains only a small subset of yeast SRB/mediator subunits (MED6, MED7, and SRB7) but a vast majority of the molecularly characterized com-

Table 5. Summary of the gene predictions in *Drosophila*. Gene prediction programs were used in combination with searches of protein and EST databases.

Result	Genie + Genscan*	Genie only†	Genscan only‡	No gene prediction§	Total
EST + protein match	6,040	288	239	49	6,616
EST match only	1,357	143	107	34	1,641
Protein match only	2,541	157	220	78	2,996
No match	1,980	307	0	0	2,348
Total	11,918	895	627	161	13,601

*Genie and Genscan matches overlapped but were not necessarily identical. †Genie predictions in regions not predicted by Genscan. ‡Genscan predictions in regions not predicted by Genie; in the absence of database matches, >4000 Genscan predictions were not included in the annotated gene set. §Gene structures defined based on database matches in the absence of gene predictions.

ponents of mammalian coactivator complexes such as ARC/DRIP/TRAP.

Gene regulation. On the basis of similarity to known proteins, *Drosophila* appears to encode about 700 transcription factors, about half of which are zinc-finger proteins. By contrast, the worm has about 500 transcription factors, fewer than one-third of which are zinc-finger proteins (*29*). Two additional classes play key roles in regulation: the homeodomain-containing and nuclear hormone receptor–type transcription factors.

Homeodomain-containing proteins control a wide variety of developmental processes. Twenty-two new homeodomain-containing proteins were uncovered in our analysis, bringing the total to more than 100. Ten of these were members of the paired-box PRX superclass (*48*), some with known vertebrate homologs: short stature homeobox 2 (SHOX), cartilage homeoprotein 1 (CART), and the two retina-specific proteins (VSX-1 and VSX-2) of goldfish. New members were also found in the LIM and TGIF class. The two new LIM members contain a homeobox and two copies of the LIM motif; the two new TGIF members occur as a local tandem duplication on the right arm of chromosome 2. We also found single new members of the NK-2, muscle-specific homeobox, proline-rich homeodomain (PRH), and BarH classes. The new fly gene encoding NK-2 is a cognate of the gene encoding the NKX-5.1 mouse protein. The new fly gene encoding muscle-specific homeobox is most similar to the gene encoding the MSX-1 mouse protein involved in craniofacial morphogenesis. The new fly gene encoding PRH is most similar to a mouse gene expressed in myeloid cells. The remaining homeodomain-containing proteins are orphans: One has similarity to the human H6 protein involved in craniofacial development, and another to HB9, a protein required for normal development of the pancreas.

Nuclear hormone receptors (NRs) are

Table 6. Gene Ontology (GO) classification of *Drosophila* gene products. Each of the 14,113 predicted transcripts was searched by BLAST against a database of proteins from fly, yeast, and mouse that had been assigned manually to a function and/or process category in the GO system. Function categories were reviewed manually, and in many cases a *Drosophila* protein was assigned to a different category upon careful inspection. The number of transcripts assigned to each process category is the result of computational searches only. For functions, the number of transcripts assigned and manually reviewed in each category is shown (with the results of the computational search in parentheses). Certain cases illustrate the value of the manual inspection. For example, motor proteins initially included many coiled-coil domain proteins incorrectly assigned to this category by the computational search. Supplemental data are available at www.celera.com.

Function	Number of transcripts
Nucleic acid binding	1387 (1370)
DNA binding	919 (652)
DNA repair protein	65 (30)
DNA replication factor	38 (18)
Transcription factor	694 (418)
RNA binding	259 (205)
Ribosomal protein	128 (116)
Translation factor	69 (68)
Transcription factor binding	21 (116)
Cell cycle regulator	52 (104)
Chaperone	159 (158)
Motor protein	98 (373)
Actin binding	93 (64)
Defense/immunity protein	47 (41)
Enzyme	2422 (2021)
Peptidase	468 (456)
Endopeptidase	378 (387)
Protein kinase	236 (307)
Protein phosphatase	93 (93)
Enzyme activator	9 (19)
Enzyme inhibitor	68 (92)
Apoptosis inhibitor	15 (17)
Signal transduction	622 (554)
Receptor	337 (336)
Transmembrane receptor	261 (280)
G protein–linked receptor	163 (160)
Olfactory receptor	48 (49)
Storage protein	12 (27)
Cell adhesion	216 (271)
Structural protein	303 (302)
Cytoskeletal structural protein	106 (54)
Transporter	665 (517)
Ion channel	148 (188)
Neurotransmitter transporter	33 (18)
Ligand binding or carrier	327 (391)
Electron transfer	124 (117)
Cytochrome P450	88 (84)
Ubiquitin	11 (17)
Tumor suppressor	10 (5)
Function unknown/unclassified	7576 (7654)
Conserved hypothetical	(1474)

Process	Number of transcripts
Cell growth and maintenance	3894
Metabolism	2274
Carbohydrate metabolism	53
Energy pathways	69
Electron transport	8
Nucleotide and nucleic acid metabolism	1078
DNA metabolism	64
DNA replication	57
DNA repair	110
DNA packaging	112
Transcription	735
Amino acid and derivative metabolism	69
Protein metabolism	685
Protein biosynthesis	215
Protein folding	52
Protein modification	273
Proteolysis and peptidolysis	81
Protein targeting	51
Lipid metabolism	111
Monocarbon compound metabolism	6
Coenzymes and prosthetic group metabolism	23
Transport	336
Ion transport	72
Small molecule transport	109
Mitochondrial transport	43
Ion homeostasis	8
Intracellular protein traffic	116
Cell death	50
Cell motility	9
Stress response	223
Defense (immune) response	149
Organelle organization and biogenesis	417
Mitochondrion organization and biogenesis	5
Cytoskeleton organization and biogenesis	390
Cytoplasm organization and biogenesis	7
Cell cycle	211
Cell communication	530
Cell adhesion	228
Signal transduction	279
Developmental processes	486
Sex determination	7
Physiological processes	201
Sensory perception	64
Behavior	54
Process unknown/unclassified	8884

sequence-specific, ligand-dependent transcription factors that contribute to physiological homeostasis by functioning as both transcriptional activators and repressors. Examination of the fly genome revealed only four additional NR members, bringing the total to 20. In contrast, the NR family represents the most abundant class of transcriptional regulators in the worm: More than 200 member genes have been described. One of the newly identified fly NRs possesses a new P-box element (Cys-Asp-Glu-Cys-Ser-Cys-Phe-Phe-Arg-Arg), which confers DNA binding specificity, bringing to 76 the number of P-boxes identified to date in all species. A search of the *Drosophila* genome failed to identify any homologs to the mammalian p160 gene family of NR coactivator proteins. SMRTER, despite weak similarity to the mammalian corepressors SMRT and N-CoR, appears to be the only close relative in *Drosophila*.

Translation and RNA processing. Although the structure of the ribosome has been well worked out, it has become apparent that many ribosomal proteins are multifunctional and are involved in processes as disparate as DNA repair and iron-binding (*49*). There has been an enormous genetic investigation of the consequences of changes in expression level of *Drosophila* ribosomal proteins (the *Minute* phenotype) (*50*); the identification and mapping of the complete set presented here will provide the basis for in-depth dissections of their functions and disease roles.

Most genes encoding general translation factors are present in only one copy in the *Drosophila* genome, as they are in other genomes studied to date; however, we discovered six genes encoding proteins highly similar to the messenger RNA (mRNA) cap-binding protein eIF4E. These may add complexity to regulation of cap-dependent translation, which is central to cellular growth control. *Caenorhabditis elegans* has three eIF4E isoforms, which were hypothesized to be necessary because trans-spliced mRNAs possess a different cap structure than do other mRNAs (*51*); however, *Drosophila* does not have trans-spliced mRNAs. The activity of eIF4E is regulated by an inhibitor protein, 4E-BP. The *Drosophila* genome contains only a single gene encoding 4E-BP; in contrast, mammals have at least three 4E-BP isoforms but perhaps fewer eIF4E isoforms than do flies. Of the more than 200 RNA-binding proteins identified, the most frequent structural classes are RRM proteins (114), DEAD- or DExH-box helicases (58), and KH-domain proteins (31). This distribution is similar to that observed in the *C. elegans* genome. These structural motifs are sometimes found in proteins for which experimental evidence indicates a function in DNA, rather than RNA, binding. Overall, the translational machinery appears well conserved throughout the eukaryotes.

The process of nonsense-mediated decay (*52*), the accelerated decay of mRNAs that cannot be translated throughout their entire length, has been genetically characterized in yeast and *C. elegans* but not in *Drosophila*. We found homologs of UPF1/SMG-2, SMG-1, and SMG-7 in the *Drosophila* genome, indicating that this process is conserved in flies.

Of particular interest are genes for components of the minor, or U12, spliceosome (*53*). Such introns are known in mammals, *Drosophila*, and *Arabidopsis*, but not *C. elegans*. Using conservative criteria (including a perfect match to the U12 consensus 5′ splice site for nucleotides 2 to 7, TATCCT), we found one intron that appears to be of the U12 type per 1000 genes. As expected, the minor spliceosome snRNAs U12, U4atac, and U6atac are present in the *Drosophila* genome. However, neither U11 nor the U11-associated 35-kD protein (*54*) could be identified in the sequence. It is possible that these components of the minor spliceosome are less well conserved, or that the minor spliceosome in *Drosophila* does not contain them.

Cytochrome P450. The cytochrome P450 monooxygenases (CYPs) are a large and ancient superfamily of proteins that carry out multiple reactions to enable organisms to rid themselves of foreign compounds. Human CYP2D6, for example, influences the metabolism of beta blockers, antidepressants, antipsychotics, and codeine, and insect CYPs function in the synthesis or degradation of hormones and pheromones and in the metabolism of natural and synthetic toxins, including insecticides (*55*). We found 90 P450 fly genes, of which four are pseudogenes, a figure that is comparable to the 80 CYPs of *C. elegans*. These 90 genes, some of which are clustered, are divided among 25 families, five of which are found in Lepidoptera, Coleoptera, Hymenoptera, Orthoptera, and Isoptera. However, more than half of the 90 genes belong to only two families, CYP4 and CYP6, the former family shared with vertebrates. CYP51, used in making cholesterol in animals and related molecules in plants and fungi, is absent from both the fly and worm genomes; it is well known that the fly must obtain cholesterol from its diet. A comprehensive collection of phylogenetically diverse CYP sequences is available (*56*).

Solute transport. Solute transporters contribute to the most basic properties of living systems, such as establishment of cell potential or generation of ATP; in higher eukaryotes, these proteins help mediate advanced functions such as behavior, learning, and memory. Hydropathy analyses predict that 20% of the gene products in *Drosophila* reside in cellular membranes, having four or more hydrophobic α helices (*57*). A considerable fraction of these proteins (657, or 4%) are dedicated to ion and metabolite movement. More than 80% of the annotated transporters are new to *Drosophila* and were identified by similarity to proteins characterized in other eukaryotes. The largest families are sugar permeases, mitochondrial carrier proteins, and the ATP-binding cassette (ABC) transporters, with 97, 38, and 48 genes, respectively; these families are also the most common in yeast and *C. elegans* (*29*). Also of note are three families of anion transporters that mediate flux of sulfate, inorganic phosphate, and iodide. Na^+-anion transporters, with 17 members, are particularly abundant relative to worm and yeast. Although individual members of these families have been investigated—for example, the mitochondrial carrier protein COLT required for gas-filling of the tracheal system (*58*) and the ABC transporters associated with eye pigment distribution (*59*)—the variety and number of transporters within each family are impressive. These data lay the foundation for understanding global transport processes critical to *Drosophila* physiology and development.

Metabolic processes. The biosynthetic networks of the fly are remarkably complete compared to those of many different prokaryotes and to yeast, in which key enzymes of various pathways may be missing (*60*). As in vertebrates, many fly enzymes are encoded by multiple genes. Two families are noteworthy because of their size. The triacylglycerol lipases are encoded by 31 genes and merit consideration in investigations of lipolysis and energy storage and redistribution. In addition, there are 32 genes encoding uridine diphosphate (UDP) glycosyltransferases, which participate in the production of sterol glycosides and in the biodegradation of hydrophobic compounds. Several UDP glycosyltransferase genes are highly expressed in the antennae and may have roles in olfaction. In vertebrates, these enzymes are critical to drug clearance and detoxification (*61*). A major challenge will be to determine whether the number of these proteins present in the genome is correlated with the importance and complexity of the regulatory events involved in any given enzymatic reaction.

Iron (Fe) is both essential for and toxic to for all living things, and metazoan animals use similar strategies for obtaining, transporting, storing, and excreting iron. Three findings from the analysis of the genome shed light on the underlying common mechanisms that have escaped attention in the past. First, a third ferritin gene has been found that probably encodes a subunit belonging to a cytosolic ferritin, the predominant type in vertebrates. This finding indicates that intracellular iron storage mechanisms in flies might be very similar to those in vertebrates. Subunits of the predominant secreted ferritins in insects are encoded by two highly expressed autosomal genes (*62*). Second, the dipteran transferrins studied so far appear to play antibiotic rather than iron-transport roles; one such transferrin was previously characterized in *Drosophila* (*63*). We have now identified two additional transferrins. The conservation of iron-binding residues and COOH-terminal hydrophobic sequences in these new transferrins suggests that they are homologs of the human melanotransferrin p97. The latter is anchored to the cells and mediates iron uptake independently from the main vertebrate pathway that involves serum transferrin and its receptor (*64*). Third, proteins homologous to vertebrate transferrin receptors appear to be absent from the fly. Thus, the *Drosophila* homologs of the vertebrate melanotransferrin could mediate the main insect pathway for cellular uptake of iron and possibly of other metal and nonmetal small ligands. This appears to be an ancestral mechanism, and the exploration of these findings should be crucial in bringing together what has seemed to be divergent iron homeostasis strategies in vertebrates and insects.

This initial look at the genomic basis of the fly's fundamental biochemical pathways reveals that its biosynthetic networks are fairly consistent with those of worm and human. On the other hand, there are a number of new findings. The large diversity of transcription factors, including several hundred zinc-finger proteins and novel homeodomain-containing proteins and nuclear hormone receptors, is likely related to the substantial regulatory complexity of the fly. In addition, many of the genes involved in core processes are single-copy genes and thus provide starting points for detailed studies of phenotype, free of the complications of genetically redundant relatives.

Fig. 4. Coding content of the fly genome. Each predicted gene in the genome is depicted as a box color-coded by similarity to genes from mammals, *C. elegans*, and *S. cerevisiae*. A legend appears at the end of each chromosome arm describing the components of each panel. In order from the top, they are (A) scale in megabases, (B) polytene chromosome divisions, (C) GC content in a range from 25 to 65%, (D) transposable elements, and genes on the (E) plus and (F) minus strands. The width of each gene element represents the total genomic length of the transcription unit. The height of each gene element represents EST coverage: The shortest boxes have no EST matches, medium-size boxes have 1 to 12 EST matches, and the tallest boxes have 13 or more EST matches. The color code for sequence similarity appears on each side of the fold-out figure. The graphics for this figure were prepared using gff2ps (*68*). Each gene has been assigned a FlyBase identifier (FBgn) in addition to the Celera identifier (CT#). Access to supporting information on each gene is available through FlyBase at http://flybase.bio.indiana.edu. These data are also available through a graphical viewing tool at FlyBase (http://flybase.bio.indiana.edu) and Celera (www.celera.com), with additional supporting information.

Concluding Remarks

Genome assembly relied on the use of several types of data, including clone-based sequence, whole-genome sequence from libraries with three insert sizes, and a BAC-based STS content map. The combination of these resources resulted in a set of ordered contigs spanning nearly all of the euchromatic region on each chromosome arm. We are taking advantage of the cloned DNA available from both the clone-based and whole-genome subclones to fill the gaps between contigs; 331 have been filled, and the remainder are in progress.

It is useful to consider the relative contributions of the various data types to the finished product with respect to how similar programs might be carried out in the future. The BAC end-sequences and STS content map provided the most informative long-range sequence-based information at the lowest cost. Both BAC ends and STS map were necessary to link scaffolds to chromosomal locations. A higher density of BAC end-sequences, from libraries produced with a larger diversity of restriction enzymes (or even from a random-shear library), would have resulted in larger scaffolds at lower shotgun sequence coverage; this is our primary recommendation for future projects. Although the clone-based draft sequence data did not result in a markedly different extent of scaffold coverage compared to assembly without the clone-based data, they were useful in the resolution of repeated sequences, particularly in the transition zones between euchromatin and centric heterochromatin. In terms of sequence coverage, adequate scaffold size was obtained with whole-genome sequence coverage as low as 6.5× (*11*). The assembly algorithm did not take any specific advantage of the fact that each draft sequence read from a BAC clone came from a defined region of the genome. Adding this feature could mean that adequate genome assembly could be obtained at lower whole-genome sequence coverage. Contiguity and scaffold size continued to increase with increased coverage, and so a decision to proceed with additional sequencing versus more directed gap closure should be driven by available resources.

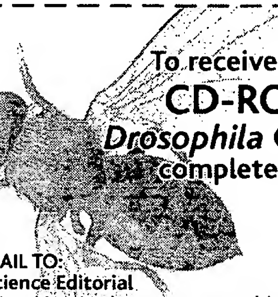
The assembled sequence has allowed a first look at the overall *Drosophila* genome structure. As previously suspected, there is no clear boundary between euchromatin and heterochromatin. Rather, over a region of ~ 1 Mb, there is a gradual increase in the density of transposable elements and other repeats, to the point that the sequence is nearly all repetitive. However, there are clearly genes within heterochromatin, and we suspect that most of our 3.8 Mb of unmapped scaffolds represent such genes, both near the centromeres and on the Y chromosome (which is almost entirely heterochromatic). Access to these sequences was an unexpected benefit of the WGS approach.

The genome sequence and the set of 13,601 predicted genes presented here are considered Release 1. Both will evolve over time as additional sequence gaps are closed, annotations are improved, cDNAs are sequenced, and genes are functionally characterized. The diversity of predicted genes and gene products will serve as the raw material for continued experimental work aimed at unraveling the molecular mechanisms underlying development, behavior, aging, and many other processes common to metazoans for which *Drosophila* is such an excellent model.

References and Notes

1. G. L. G. Miklos and G. M. Rubin, *Cell* **86**, 521 (1996).
2. A. S. Spradling *et al.*, *Genetics* **153**, 135 (1999).
3. M. Ashburner *et al.*, *Genetics* **153**, 179 (1999).
4. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
5. G. M. Rubin and E. Lewis, *Science* **287**, 2216 (2000).
6. D. L. Hartl *et al.*, *Trends Genet.* **8**, 70 (1992).
7. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser and R. D. Fleischmann, *Electrophoresis* **18**, 1207 (1997).
8. J. L. Weber and E. W. Myers, *Genome Res.* **7**, 409 (1997).
9. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
10. R. Hoskins *et al.*, *Science* **287**, 2271 (2000).
11. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
12. A number of methods were used to close gaps. Whenever possible, gaps were localized to a chromosome region and a spanning genomic clone was identified. When a spanning clone could be identified, it was used as a template for sequencing. The sequencing approach was determined by the gap size. For gaps smaller than 1 kb, BAC templates were sequenced directly with custom primers. For gaps larger than 1 kb, 3-kb plasmids or M13 clones from the clone-based draft sequencing were sequenced by directed methods, or 10-kb plasmids from the WGS sequencing project were sequenced by random transposon-based methods. If no 3-kb or 10-kb plasmid could be identified, PCR products were amplified from BAC clones or genomic DNA and end-sequenced directly with the PCR primers.
13. K. S. Weiler and B. T. Wakimoto, *Annu. Rev. Genet.* **29**, 577 (1995); S. Henikoff, *Biochem. Biophys. Acta* **1470**, 1 (2000); S. Pimpinelli *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3804 (1995); A. R. Lohe, A. J. Hilliker, P. A. Roberts, *Genetics* **134**, 1149 (1993).
14. G. L. G. Miklos, M. Yamamoto, J. Davies, V. Pirrotta, *Proc. Natl. Acad. Sci U.S.A.* **85**, 2051 (1988).
15. See ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/nuclear_cds_set.embl.v2.9.Z.
16. The genes found in unscaffolded sequence were Su(Ste) (FlyBase identifier FBgn0003582) on the Y chromosome, His1 (FBgn0001195) and His4 (FBgn0001200) (histone genes were screened out before assembly), rbp13 (FBgn0014016), and idr (FBgn0020850).
17. C. Burge and S. Karlin, *J. Mol. Biol.* **268**, 78 (1997).
18. M. G. Reese, D. Kulp, H. Tammana, D. Haussler, *Genome Res.*, in press.

19. Sequence contigs were searched against publicly available sequence at the DNA level and as six-frame translations against public protein sequence data. DNA searches were against the invertebrate (INV) division of GenBank, a set of 80,000 EST sequences produced at BDGP assembled to produce consensus sequences (*21*), and a set of curated *Drosophila* protein-coding genes prepared by three of the authors (M. Ashburner, L. Bayraktaroglu, and P. V. Benos) (*15*). Protein searches were performed against this set of curated protein sequences and against the nonredundant protein database available at the National Center for Biotechnology Information. Initial searches were performed with a version of BLAST2 (*25*), optimized for the Compaq Alpha architecture. Additional processing of each query-subject pair was performed to improve the alignments. All BLAST results having an expectation score of $<1 \times 10^{-4}$ were then processed on the basis of their high-scoring pair (HSP) coordinates on the contig to remove redundant hits, retaining hits that supported possible alternative splicing. This procedure was performed separately by hits to particular organisms so as not to exclude HSPs that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the contig sequence were then realigned to the contig with Sim4 [L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, W. Miller, *Genome Res.* **8**, 967 (1998)] for ESTs, and with Lap [X. Huang, M. D. Adams, H. Zhou, A. R. Kerlavage, *Genomics* **46**, 37 (1995)] for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually respect intron-exon boundaries and thus facilitate human annotation. Some regions of the genome may be underannotated because the bulk of the annotation work was done on an earlier assembly version. Continued updates will be available through FlyBase.
20. M. G. Reese, G. Hartzell, N. L. Harris, U. Ohler, S. E. Lewis, *Genome Res.*, in press.
21. G. M. Rubin *et al.*, *Science* **287**, 2222 (2000).
22. See the Gene Ontology Web site (www.geneontology.org).
23. See the *Saccharomyces* Genome Database Web site (http://genome-www.stanford.edu/Saccharomyces).
24. D. Allen and J. Blake, Mouse Genome Informatics (www.informatics.jax.org).
25. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
26. S. M. Mount *et al.*, *Nucleic Acids Res.* **20**, 4255 (1992).
27. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
28. X. Lin *et al.*, *Nature* **402**, 761 (1999).
29. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
30. A. Dutta and S. P. Bell, *Annu. Rev. Cell Dev. Biol.* **13**, 293 (1997).
31. I. Chesnokov, M. Gossen, D. Remus, M. Botchan, *Genes Dev.* **13**, 1288 (1999).
32. G. Feger, *Gene* **227**, 149 (1999).
33. D. T. Pak *et al.*, *Cell* **97**, 311 (1997); J. Rohrbough, S. Pinto, R. M. Mihalek, T. Tully, K. Broadie, *Neuron* **23**, 55 (1999).
34. S. Waga, G. J. Hannon, D. Beach, B. Stillman, *Nature* **369**, 574 (1994); H. Flores-Rozas *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 8655 (1994).
35. R. Jessberger, C. Frei, S. M. Gasser, *Curr. Opin. Genet. Dev.* **8**, 254 (1998); T. Hirano, *Curr. Opin. Genet. Dev.* **10**, 317 (1998); A. V. Strunnikov, *Trends Cell Biol.* **8**, 454 (1998).
36. R. Saffery *et al.*, *Hum. Mol. Genet.* **9**, 175 (2000); J. M. Craig, W. C. Earnshaw, P. Vagnarelli, *Exp. Cell Res.* **246**, 249 (1999); R. Saffery *et al.*, *Chromosome Res.* **7**, 261 (1996).
37. R. Belotserkovskaya and S. L. Berger, *Crit. Rev. Eukaryotic Gene Expr.* **9**, 221 (1999).
38. J. A. Eisen, K. S. Sweder, P. C. Hanawalt, *Nucleic Acids Res.* **23**, 2715 (1995); K. J. Pollard and C. L. Peterson, *Bioessays* **20**, 771 (1998).
39. E. V. Koonin, S. Zhou, J. C. Lucchesi, *Nucleic Acids Res.* **23**, 4229 (1995).
40. F. Jeanmougin *et al.*, *Trends Biochem. Sci.* **22**, 151 (1997); F. Winston and C. D. Allis, *Nature Struct. Biol.* **6**, 601 (1999).
41. R. W. Levis, *Mol. Gen. Genet.* **236**, 440 (1993); H. Biessmann and J. M. Mason, *Chromosoma* **106**, 63 (1997).
42. P. Gallinari and J. Jiricny, *Nature* **383**, 735 (1996).
43. B. Flores and W. Engels, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2964 (1999).
44. K. Kusano, M. E. Berres, W. R. Engels, *Genetics* **151**, 1027 (1999); J. J. Sekelsky, M. H. Brodsky, G. M. Rubin, R. S. Hawley, *Nucleic Acids Res.* **27**, 3762 (1999).
45. M. Hampsey, *Microbiol. Mol. Biol. Rev.* **62**, 465 (1998); R. H. Reeder, *Prog. Nucleic Acid Res. Mol. Biol.* **62**, 293 (1999); I. M. Willis, *Eur. J. Biochem.* **212**, 1 (1993).
46. T. I. Lee and R. A. Young, *Genes Dev.* **12**, 1398 (1998); M. Hampsey and D. Reinberg, *Curr. Opin. Genet. Dev.* **9**, 132 (1999).
47. M. D. Rabenstein, S. Zhou, J. T. Lis, R. Tjian, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4791 (1999).
48. D. Duboule, Ed., *Guidebook to the Homeobox Genes* (Oxford Univ. Press, New York, 1994).
49. I. G. Wool, *Trends Biochem. Sci.* **21**, 164 (1996).
50. A. Lambertsson, *Adv. Genet.* **38**, 69 (1998).
51. M. Jankowska-Anyszka *et al.*, *J. Biol. Chem.* **273**, 10538 (1998).
52. M. R. Culbertson, *Trends Genet.* **15**, 74 (1999).
53. C. Burge, T. Tuschl, P. Sharp, in *The RNA World*, R. Gesteland, T. Cech, J. Atkins, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1999).
54. C. L. Will, C. Schneider, R. Reed, R. Luhrmann, *Science* **284**, 2003 (1999).
55. R. Feyereisen, *Annu. Rev. Entomol.* **44**, 507 (1999).
56. See D. Nelson's Web site (http://drnelson.utmem.edu/CytochromeP450.html).
57. G. von Heijne, *J. Mol. Biol.* **225**, 487 (1992).
58. K. Hartenstein *et al.*, *Genetics* **147**, 1755 (1997).
59. R. G. Tearle, J. M. Belote, M. McKeown, B. S. Baker, A. J. Howells, *Genetics* **122**, 595 (1989).
60. R. Maleszka, *Microbiology* **143**, 1781 (1997).
61. Q. Wang, G. Hasan, C. W. Pikielny, *J. Biol. Chem.* **274**, 10309 (1999).
62. B. C. Dunkov and T. Georgieva, *DNA Cell Biol.* **18**, 937 (1999).
63. T. Yoshiga *et al.*, *Eur. J. Biochem.* **260**, 414 (1999).
64. M. L. Kennard *et al.*, *EMBO J.* **14**, 4178 (1995).
65. High molecular weight genomic DNA was prepared from nuclei isolated [C. D. Shaffer, J. M. Wuller, S. C. R. Elgin, *Methods Cell Biol.* **44**, 185 (1994)] from 2.59 g of embryos of an isogenic *y; cn bw sp* strain [B. J. Brizuela *et al.*, *Genetics* **137**, 803 (1994)]. The genomic DNA was randomly sheared, end-polished with Bal31 nuclease/T4 DNA polymerase, and carefully size-selected on 1% low-melting-point agarose. After ligation to BstX1 adaptors, genomic fragments were inserted into BstX1-linearized plasmid vector. Libraries of 1.8 ± 0.2 kb were cloned in a high-copy pUC18 derivative, and libraries of 9.8 ± 1.0, 10.5 ± 1.0, and 11.5 ± 1.0 kbp were cloned in a medium-copy pBR322 derivative. High-throughput methods in 384-well format were implemented for plasmid growth, alkaline lysis plasmid purification, and ABI Big Dye Terminator DNA sequencing reactions. Sequence reads from the genomic libraries were generated over a 4-month period using 300 DNA analyzers (ABI Prism 3700). These reads represent more than 12× coverage of the 120-Mbp euchromatic portion of the *Drosophila* genome (Table 1). Base-calling was performed using 3700 Data Collection (PE Biosystems) and sequence data were transferred to a Unix computer environment for further processing. Error probabilities were assigned to each base with TraceTuner software developed at Paracel Inc. (www.paracel.com). The predicted error probability was used to trim each sequence read such that the overall accuracy of each trimmed read was predicted to be >98.5% and no single 50-bp region was less than 97% accurate. The efficacy of TraceTuner and the trimming algorithm was demonstrated by comparing trimmed sequence reads to high-quality finished sequence data from BDGP (Fig. 2).
66. For clone-based genomic sequencing, BAC, P1, and cosmid DNAs were prepared by alkaline lysis procedures and purified by CsCl gradient ultracentrifugation. DNA was randomly sheared and size-selected on LMP agarose for fragments in the 3-kb range for plasmids and in the 2-kb range for M13 clones. After blunt-ending with T4 DNA polymerase, plasmids were generated by ligation to BstX1 adaptors and insertion into BstX1-linearized pOT2A vector. M13 clones were generated using the double-adaptor protocol [B. Andersson *et al.*, *Anal. Biochem.* **236**, 107 (1996)]. Plasmid sequencing templates were prepared by alkaline lysis (Qiagen) or by PCR, and M13 templates were prepared using the sodium perchlorate–glass fiber filter technique [B. Andersson *et al.*, *Biotechniques* **20**, 1022 (1996)]. Paired end-sequences of 3-kb plasmid subclones were generated (principally) with ABI Big Dye Terminator chemistry on ABI 377 slab gel or ABI 3700 capillary sequencers. Additional M13 subclone sequence was generated using BODIPY dye-labeled primers. Procedures for finishing sequence to high quality at LBNL were as described (*3*).
67. M.-T. Yamamoto *et al.*, *Genetics* **125**, 821 (1990).
68. J. F. Abril and R. Guigo, *Bioinformatics*, in press.
69. A. Peter *et al.*, in preparation.
70. J. Locke, L. Podemski, N. Aippersbach, H. Kemp, R. Hodgetts, in preparation.
71. The many participants from academic institutions are grateful for their various sources of support. We thank B. Thompson and his staff for the excellent laboratories and work environment, M. Peterson and his team for computational support, and V. Di Francesco, S. Levy, K. Chaturvedi, D. Rusch, C. Yan, and V. Bonazzi for technical discussions and thoughtful advice. We are indebted to R. Guigo and to E. Lerner of Aquent Partners for assistance with illustrations. The work described was funded by Celera Genomics, the Howard Hughes Medical Institute, and NIH grant P50-HG00750 (G.M.R.).



PROTEIN KINASES 6

The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification[1]

STEVEN K. HANKS* AND TONY HUNTER[2]

*Department of Cell Biology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA; and Molecular Biology and Virology Laboratory, The Salk Institute, San Diego, California 92186, USA

The eukaryotic protein kinases comprise one of the largest superfamilies of homologous proteins and genes. Within this family, there are now hundreds of different members whose sequences are known. Although there is a rich diversity of structures, regulation modes, and substrate specificities among the protein kinases, there are also common structural features. These conserved structural motifs provide clear indications as to how these enzymes manage to transfer the γ-phosphate of a purine nucleotide triphosphate to the hydroxyl groups of their protein substrates. The authors of this review have carried out a monumental task of analyzing and collating the amino acid sequences of all reported protein kinases and defining the conserved structural features that characterize the portion of these proteins that is responsible for their catalytic activity. Comparison of the sequences in the catalytic fragment of the protein kinases has been used to arrange these enzymes in evolutionary trees that group subfamilies of closely related enzymes. It is comforting that the structural relationships that emerge from these trees result in groupings that also reflect related functions. The work presented in this review seems to be an excellent example of the type of analysis that will become indispensable in the coming years, as more and more sequence information become available to biologists as a result of the genome projects.

ABSTRACT The eukaryotic protein kinases make up a large superfamily of homologous proteins. They are related by virtue of their kinase domains (also known as catalytic domains), which consist of ≈250–300 amino acid residues. The kinase domains that define this group of enzymes contain 12 conserved subdomains that fold into a common catalytic core structure, as revealed by the 3-dimensional structures of several protein-serine kinases. There are two main subdivisions within the superfamily: the protein-serine/threonine kinases and the protein-tyrosine kinases. A classification scheme can be founded on a kinase domain phylogeny, which reveals families of enzymes that have related substrate specificities and modes of regulation.—Hanks, S. K., Hunter, T. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. 9, 576–596 (1995)

Key Words: protein-tyrosine kinase • protein-serine kinase • protein phosphorylation • AMP-dependent protein kinase

THE EUKARYOTIC PROTEIN KINASE SUPERFAMILY

One of the largest known protein superfamilies is made up of protein kinases identified largely from eukaryotic sources. (The term superfamily will be used here to distinguish this broad collection of enzymes from smaller, more closely related subsets that have been commonly referred to as families). These enzymes use the γ-phosphate of ATP (or GTP) to generate phosphate monoesters using protein alcohol groups (on Ser and Thr) and/or protein phenolic groups (on Tyr) as phosphate acceptors. The protein kinases are related by virtue of their homologous kinase domains (also known as catalytic domains), which consist of ~250–300 amino acid residues (reviewed in refs 1–3; and see below). During the past 15 years, previously unrecognized members of the eukaryotic protein kinase superfamily have been uncovered at an exponentially increasing rate and currently appear in the literature almost weekly. This pace of discovery can be attributed to the past development of molecular cloning and sequencing technologies and, more recently, to the advent of the polymerase chain reaction (PCR),[3] which facilitated the use of homology-based cloning strategies. Consequently, about 200 different superfamily members (products of distinct paralogous genes) had been recognized from mammalian sources alone! The prediction made several years ago (4) that the mammalian genome contains about 1000 protein kinase genes (roughly 1% of all genes) would still appear to be within reason, and may even be an underestimate (5).

In addition to mammals and other vertebrates, eukaryotic protein kinase superfamily members have been identified and characterized from a wide range of other animal phyla as well as from plants, fungi, and protozoans. Hence, the protein kinase progenitor gene can be traced back to a time before the evolutionary separation of the major eukaryotic kingdoms. The identification of eukaryotic-like protein kinase genes in prokaryotes (6, 7) raises the possibility that the protein kinase progenitor gene might have arisen before the divergence of prokaryotes and eukaryotes (see below). Studies of the budding and fission yeasts, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, have been particularly fruitful in the recognition of new protein kinases. In these geneti-

[1] This article is based on an introductory chapter in the *Protein Kinase Factsbook*, edited by D. G. Hardie and S. K. Hanks, published in 1995 by Academic Press, London.

[2] To whom correspondence and reprint requests should be addressed, at: Molecular Biology and Virology Laboratory, The Salk Institute, 10010 N. Torrey Pines Rd., La Jolla, CA 92037, USA.

[3] Abbreviations: PCR, polymerase chain reaction; PKA-Cα, type α cAMP-dependent protein kinase catalytic subunit; Cdk2, cyclin-dependent kinase 2; Erk2, p42 MAP kinase; APE,

0892-6638/95/0009-0576/$01.50. © FASEB

cally tractable organisms, the powerful approach of mutant isolation and cloning by complementation has netted dozens of protein kinase genes required for numerous aspects of cell function (8). In many cases, vertebrate counterparts have now been found for these genes, leading to a growing awareness that protein phosphorylation pathways that regulate basic aspects of cell physiology have been maintained throughout the course of eukaryotic evolution.

Even though the overwhelming majority of protein kinases identified from eukaryotic sources belong to this superfamily, a small but growing number of such enzymes do not qualify as superfamily members. Most of these are related to the prokaryotic protein-histidine kinase family (see below), which forms the sensor components of two-component signal transduction systems (9). Included in this category are a putative ethylene receptor encoded by the flowering plant *ETR1* gene (10), the product of the budding yeast *SLN1* gene (11, 12) thought to be involved in relaying nutrient information to elements controlling cell growth and division, the mitochondrial branched-chain α-ketoacid dehydrogenase kinase (13), and the mitochondrial pyruvate dehydrogenase kinase (14). In prokaryotes, protein-histidine kinases phosphorylate aspartates in their target proteins, but except for the two dehydrogenase kinases that phosphorylate serine, the acceptor specificities of most of the eukaryotic protein kinases of this type are not known. In addition to these protein kinases, the Bcr protein encoded by the *breakpoint cluster region* gene involved in the Philadelphia chromosome translocation (15) and the A6 kinase isolated by expression cloning using an anti-phosphotyrosine antibody (16) have kinase domains unrelated to any known eukaryotic or prokaryotic kinase. In addition, true protein-histidine kinases are known in eukaryotes. One such enzyme has been extensively characterized from budding yeast but not yet molecularly cloned (17), and so it is not clear whether this enzyme will belong to the protein kinase superfamily or use a novel structural principle for phosphotransfer.

What about the prokaryotes? It has been known for years that protein phosphorylation events play key regulatory roles in numerous bacterial cell processes including chemotaxis, bacteriophage infection, nutrient uptake, and gene transcription (reviewed in refs 18, 19). The bacterial protein kinases have been divided into three general classes (20): *1)* protein-histidine kinases such as those functioning in two-component sensory regulatory systems (strictly speaking, these are protein-aspartyl kinases, because autophosphorylation on His is an intermediary step in phosphotransfer to an aspartate in the response-regulator protein) (9); *2)* phosphotransferases such as those of the phosphoenol pyruvate-dependent phosphotransferase system involved in sugar uptake (21); and *3)* protein-serine kinases such as isocitrate dehydrogenase kinase/phosphatase (22). Amino acid sequences have been determined for members of each class, and all are unrelated to the eukaryotic protein kinase superfamily.

Recently, however, true homologs of the eukaryotic protein kinases have been identified from two species of bacteria, *Yersinia pseudotuberculosis* (7) and *Myxococcus xanthus* (6, 23). Are these special cases, or the first examples of many such genes in prokaryotes? The eukaryotic-like protein kinase YpkA from the pathogenic enterobacteria *Y. pseudotuberculosis* is encoded by a plasmid essential for the virulence of this infectious organism. In addition to YpkA, at least two other proteins encoded by genes residing on the virulence plasmid exhibit high similarity to eukaryotic proteins. Thus, it seems likely that the virulence plasmid genes were transduced from a eukaryotic host by horizontal transfer. The myxobacterium *M. xanthus* presents a different and perhaps more intriguing picture. Application of the PCR homology-based cloning strategy revealed that at least eight genes encoding members of the eukaryotic protein kinase superfamily are present in the genome of this species (23). The myxobacteria are unusual prokaryotes in that they undergo a complex developmental cycle upon nutrient depletion, much like that of the eukaryotic slime mold *Dictyostelium*. Given that protein kinases are commonly involved in regulating growth and differentiation of eukaryotic cells, it is attractive to speculate that the eukaryotic-like protein kinases in *M. xanthus* are specifically involved in regulating their developmental cycle. Indeed, one of these kinases, Pkn1, was shown to be required for proper fruiting body formation. The same could be true for the eukaryotic-like protein kinase PknA from *Anabena* (24). In keeping with this idea, neither the PCR approach applied to *Escherichia coli* (23) nor extensive sequencing of the *E. coli* genome (now 30% complete) has yielded eukaryotic-like protein kinases. Hence, genes encoding members of the eukaryotic protein kinase superfamily may be present only in bacteria that can undergo a developmental cycle. However, unpublished reports of eukaryotic-like protein kinases in *Streptomyces coelicolor*, and in three species of *Methanococcus*, suggest that such genes are more widely expressed among prokaryotes, and potentially these genes represent the ancestors for the entire eukaryotic protein kinase superfamily.

THE HOMOLOGOUS KINASE DOMAINS

The kinase domains of eukaryotic protein kinases impart the catalytic activity. Three separate roles can be ascribed to the kinase domains: *1)* binding and orientation of the ATP (or GTP) phosphate donor as a complex with divalent cation (usually Mg^{2+} or Mn^{2+}); *2)* binding and orientation of the protein (or peptide) substrate; and *3)* transfer of the γ-phosphate from ATP (or GTP) to the acceptor hydroxyl residue (Ser, Thr, or Tyr) of the protein substrate.

Conserved features of primary structure

The total number of distinct kinase domain amino acid sequences available is now approaching 400 (**Table 1**). Included in this total are the vertebrate enzymes encoded by distinct paralogous genes, their presumed functional homologs from invertebrates and simpler organisms (encoded by orthologous genes), and those identified from lower organisms and plants for which vertebrate equivalents have not been found. Conserved features of kinase domain primary structure have previously been identified through an inspection of multiple amino acid sequence alignments (1-3) . The large number of sequences now available precludes showing an alignment containing all known kinase domains. Thus, in **Fig. 1** only 60 different kinase domain sequences are aligned. These are drawn, however, from the widest possible sampling of the superfamily and thus provide a good representation of the

Table 1. *Eukaryotic protein kinase superfamily classification.*

A-C-G Group	
AGC-I. Cyclic nucleotide-regulated protein kinase family	
A. Cyclic AMP-dependent protein kinase (PKA) subfamily	
vertebrate:	
1. PKA-Cα:	PKA catalytic subunit, alpha-form
2. PKA-Cβ:	PKA catalytic subunit, beta-form
3. PKA-Cγ:	PKA catalytic subunit, gamma-form
Drosophila melanogaster:	
1. DmPKA-C0:	PKA catalytic subunit, C0 form
2. DmPKA-C1:	PKA catalytic subunit, C1 form
3. DmPKA-C2:	PKA catalytic subunit, C2 form
Caenorhabditis elegans:	
1. CePKA:	PKA catalytic subunit homolog
Saccharomyces cerevisiae:	
1. ScPKA-Tpk1:	PKA catalytic subunit homolog, type 1
Schizosaccharomyces pombe:	
1. SpPKA1:	PKA catalytic subunit homolog
Dictyostelium discoideum:	
1. DdPKA:	PKA catalytic subunit
Aplysia californica:	
1. AplC:	PKA catalytic subunit homolog
2. Sak:	"Spermatozoon-associated kinase"
B. Cyclic GMP-dependent protein kinase (PKG) subfamily	
vertebrate:	
1. PKG-I:	PKG, type I
* 2. PKG-II:	PKG, type II
Drosophila melanogaster:	
1. DmPKG-G1:	PKG homolog, type 1
2. DmPKG-G2:	PKG homolog, type 2
C. Others	
Dictyostelium discoideum:	
1. DdPK1:	PKA homolog
AGC-II. Diacylglycerol-activated/phospholipid-dependent protein kinase C (PKC) family	
A. "Conventional" (Ca^{2+}-dependent) protein kinase C (cPKC) subfamily	
vertebrate:	
1. cPKCα:	Protein Kinase C, alpha-form
2. cPKCβ:	Protein Kinase C, beta-form
3. cPKCγ:	Protein Kinase C, gamma-form
Drosophila melanogaster:	
1. DmPKC-53Ebr:	PKC homolog expressed in brain, locus 53E
2. DmPKC-53Eey:	PKC homolog expressed in eye, locus 53E
Aplysia californica:	
1. Apl-I:	PKC homolog, type I
B. "Novel" (Ca^{2+}-independent) Protein Kinase C (nPKC) subfamily	
vertebrate:	
1. nPKCδ:	Protein Kinase C, delta-form
2. nPKCε:	Protein Kinase C, epsilon-form
3. nPKCη:	Protein Kinase C, eta-form
4. nPKCθ:	Protein Kinase C, theta-form
Drosophila melanogaster:	
1. DmPKC-98F:	PKC homolog, locus 98F
Aplysia californica:	
1. Apl-II:	PKC homolog, type II
Caenorhabditis elegans:	
1. CePKC:	PKC homolog, product of *tpa-1* gene
* 2. CePKC1B:	PKC homolog expressed in neurons and interneurons
Dictyostelium discoideum:	
* 1. DdMHCK:	PKC homolog
Saccharomyces cerevisiae:	
1. ScPKA1:	PKC homolog, product of *PKC1* gene
* 2. ScPKA2:	PKC homolog, product of *PKC2* gene
Schizosaccharomyces pombe:	
1. Pck1:	"Pombe C-kinase", type 1
2. Pck2:	"Pombe C-kinase", type 2
C. "Atypical" Protein Kinase C (aPKC) subfamily	
vertebrate:	
1. aPKCζ:	Protein Kinase C, zeta-form
* 2. aPKCι:	Protein Kinase C, iota-form
* 4. aPKCμ:	Protein Kinase C, mu-form

[a] More information about the individual protein kinases listed (including sequence references) can be obtained by contacting the authors or by consulting *The Protein Kinase Factsbook* (42). Protein kinases marked with asterisks (*) were not included in the phylogenetic analysis due to their recent discovery. In many instances new protein kinases were cloned by more than one group; in these cases the most commonly accepted name is used for the entry and alternative names are listed in parentheses after the entry. Protein kinase homologs from DNA viruses are not included in this classification.

Table 1. *(continued).*

	D. Others	
	vertebrate:	
*	1. PKN:	Protein kinase with PKC-related catalytic domain
	AGC-III. Related to PKA and PKC (RAC) family	
	vertebrate:	
	1. RAC-α:	RAC, alpha-form; cellular homolog of v-Akt oncoprotein
	2. RAC-β:	RAC, beta-form
	Drosophila:	
	1. DmRAC:	RAC homolog
	Caenorhabditis elegans:	
*	1. CeRAC:	RAC homolog
	ACG-IV. Family of kinasese that phosphorylate G protein-coupled receptors	
	vertebrate:	
	1. βARK1:	β-adrenergic receptor kinase, type 1
	2. βARK2:	β-adrenergic receptor kinase, type 2
	3.RhK:	Rhodopsin kinase
*	4.IT11:	G-protein-coupled receptor kinase homolog
*	5.GRK5:	G-protein-coupled receptor kinase, type 5
*	6. GRK6:	G-protein-coupled receptor kinase, type 6
	Drosophila melanogaster:	
	1. DmGPRK1:	Drosophila G-protein-coupled receptor kinase, type 1
	2. DmGPRK2:	Drosophila G-protein-coupled receptor kinase, type 2
	AGC-V. Family of budding yeast AGC-related kinases	
	Saccharomyces cerevisiae:	
	1. Sch9:	Suppressor of defects in cAMP effector pathway
	2. Ykr2:	AGC-related kinase
	3. Ypk1:`	AGC-related kinase
	AGC-VI. Family of kinases that phosphorylate ribosomal S6 protein	
	vertebrate:	
	1. S6K:	70 kDa S6 kinase with single catalytic domain
	2. RSK1(Nt):	90 kDA S6 kinase, type 1
	3. RSK2(Nt):	90 kDA S6 kinase, type 2
	[Note: The RSK enzymes have two distinct catalytic domains. The Nt-domain is closely related to S6K, whereas the Ct-domain is most closely related to phosphorylase kinase]	
	AGC-VII. Budding yeast Dbf2/20 Family	
	Saccharomyces cerevisiae:	
	1. Dbf2:	Product of gene periodically expressed in cell cycle
	2. Dbf20:	Close relative of DBF2 not under cell cycle control
	AG-VIII. Flowering plant "PVPK1 Family" of protein kinase homologs	
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. PvK1:	Bean protein kinase homolog
	2. OsG11A:	Rice protein kinase homolog
	3. ZmPPK:	Maize protein kinase homolog
	4. AtPK5:	Arabidopsis protein kinase homolog
	5. AtPK7:	Arabidopsis protein kinase homolog
	6. AtPK64:	Arabidopsis protein kinase homolog
	7. PsPK5:	Pea protein kinase homolog
	Other AGC-related kinases	
	vertebrate:	
	1. DMPK:	"Myotonic Dystrophy Protein Kinase"
	2. Sgk:	"Serum and glucocortocoid regulated kinase"
*	3. Mast205:	Spermatid "Microtubule-associated serine/threonine kinase"
	Neurospora crassa:	
	1. NcCot-1:	Product of gene required for normal colonial growth
	Dictyostelium discoideum:	
	1. Ddk2:	Product of developmentally-regulated gene
	Saccharomyces cerevisiae:	
	1. ScSpk1:	Dual-specificity kinase
	Phylum Angiospermophyta (Kingdom Plantae):	
*	1. Atpk1:	Arabidopsis protein kinase
	CaMK Group	
	CaMK-I. Family of kinases regulated by Ca^{2+}/Calmodulin, and close relatives	
	A. Subfamily including "Multifunctional" Ca^{2+}/Calmodulin Kinases (CaMKs)	
	vertebrate:	
	1. CaMK1:	CaMK, type I
	2. CaMK2α:	CaMK, type II, alpha subunit
	3. CaMK2β:	CaMK, type II, beta subunit
	4. CaMK2γ:	CaMK, type II, gamma subunit
	5. CaMK2δ:	CaMK, type II, delta subunit
*	6. EF2K:	Elongation Factor-2 Kinase or CaMK type III
	7. CaMK4:	CaMK, type IV

Table 1. *(continued).*

	Drosophila melanogaster:	
	1. DmCaMK2:	CaMK-II homolog
	Saccharomyces cerevisiae:	
	1. ScCaMK2-1:	CaMK-II homolog, product of *CMK1* gene
	2. ScCaMK2-2:	CaMK-II homolog, product of *CMK2* gene
	Aspergillus nidulans:	
	1. AnCaMK2:	CaMK-II homolog
	B. Subfamily including phosphorylase kinases	
	vertebrate:	
	1. PhK-γM:	Skeletal muscle phosphorylase kinase catalytic subunit
	2. PhK-γT:	Male germ cell phosphorylase kinase catalytic subunit
	3. RSK1(Ct):	90 kDa S6 kinase, type 1; C-terminal catalytic domain
	4. RSK2(Ct):	90 kDa S6 kinase, type 2; C-terminal catalytic domain
	C. Subfamily including myosin light chain kinases	
	vertebrate:	
	1. skMLCK:	Skeletal muscle MLCK (rabbit)
	2. smMLCK:	Smooth muscle MLCK (rabbit)
	3. Titin:	Huge protein implicated in skeletal muscle development
	Caenorhabditis elegans:	
	1. Twn:	"Twitchin" protein involved in muscle contraction or development
	Dictyostelium discoideum:	
	1. DdMLCK:	Slime mold myosin light chain kinase
	D. Subfamily of plant kinases with intrinsic calmodulin-like domain	
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. CDPK:	Soybean Ca^{2+}-regulated kinase with intrinsic CaM-like domain
	2. AtAK1:	Arabidopsis CDPK homolog
*	3. OsSpk:	Rice CDPK homolog
*	4. DcPk431:	Carrot CDPK homolog
	E. Subfamily of plant kinases with highly acidic domain	
	Phylum Angiospermophyta (Kingdom Plantae):	
*	1. ASK1:	Arabidopsis protein kinase homolog with highly acidic idomain
*	2. ASK2:	Arabidopsis protein kinase homolog with highly acidic domain
	F. Other CaMK-related kinases	
	vertebrate:	
	1. PskH1:	Putative protein-serine kinase
*	2. MAPKAP2:	"MAP Kinase-Activated Protein Kinase 2"
	Saccharomyces cerevisiae:	
	1. Mre4:	Protein required for meiotic recombination
*	2. Dun1:	Protein required for DNA damage-inducible gene expression
*	3. Rck1:	"Radiation sensitivity complementing kinase, type 1"
*	4. Rck2:	"Radiation sensitivity complementing kinase, type 2"
	CaMK-II. Snfl/AMPK family	
	vertebrate:	
*	1: AMPK:	"AMP-Activated Protein Kinase"
	2: p78:	Protein lost in carcinomas of human pancreas
	Saccharomyces cerevisiae:	
	1. Snfl:	Kinase essential for release from glucose repression
	2. Kin1:	Protein kinase with N-terminal catalytic domain
	3. Kin2:	Close relative of KIN1
	4. Ycl24:	Protein kinase homolog on chromosome III
*	5. Ycl453:	Protein kinase homolog on chromosome XI
	Schizosaccharomyces pombe:	
	1. SpKin1:	Product of gene important for growth polarity
	2. Nim1:	Inducer of mitosis
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. PSnfl-RKIN1:	Rye putative protein kinase that complements yeast *snf1* polarity
	2. PSnfl-AKIN10:	Arabidopsis putative protein kinase related to SNF1
	3. PSnfl-BKIN12:	Barley protein related to SNF1
*	4. PKABA1:	Wheat kinase induced by abscisic acid
*	5. WPK4:	Wheat kinase homolog regulated by light and nutrients
*	6. NPK5:	Tobacco Snfl homolog, activates *SUC2* gene expression
	Other CaMK Group Kinases	
	Plasmodium falciparum (malarial parasite):	
	1. PfCPK:	Ca^{2+}-regulated kinase with intrinsic CaM-like domain
	2. PfPK2:	Putative protein kinase
	C-M-G-C Group	
	CMGC-I. Family of cyclin-dependent kinases (CDKs) and other close relatives	
	vertebrate:	
	1. Cdc2:	Inducer of mitosis; functional homolog of yeast cdc2+/CDC28 kinases (Cdk1)
	2. Cdk2:	Type 2 cyclin-dependent kinase
	3. Cdk3:	Type 3 cyclin-dependent kinase
	4. Cdk4:	Type 4 cyclin-dependent kinase
	5. Cdk5:	Type 5 cyclin-dependent kinase

Table 1. *(continued)*.

	6. Cdk6:	Type 6 cyclin-dependent kinase
	7. PCTAIRE1:	Cdc2-related protein
	8. PCTAIRE2:	Cdc2-related protein
	9. PCTAIRE3:	Cdc2-related protein
	10. Mo15:	"Cdk-activating kinase"; Negative regulator of meiosis (CAK)
	Drosophila melanogaster:	
	1. DmCdc2:	Functional homolog of yeast cdc2+/CDC28 kinases
	2. DmCdc2c:	Cdc2-cognate protein; Cdk2 homolog
	Dictyostelium discoideum:	
	1. DdCdc2:	Functional homolog of yeast cdc2+/CDC28 kinases
	2. DdPRK:	"Cdc2-related PCTAIRE Kinase"
	Aspergillus nidulans:	
	1. NIMXcdc2:	Cdc2-related gene product
	Plasmodium falciparum:	
	1. PfPK5:	Cdc2-related protein from human malarial parasite
	Entamoeba histolytica:	
	1. EhC2R:	Cdc2-related protein
	Crithidia fasciculata:	
	1. CfCdc2R:	Cdc2-related protein
	Leishmania mexicana:	
*	1. LmCRK1:	"Cdc2-Related Kinase"
	Saccharomyces cerevisiae:	
	1. Cdc28:	"Cell-division-cycle" gene product
	2. Pho85:	Negative regulator of the PHO system and cell cycle regulator
	3. Kin28:	CDC28-related protein
	Schizosaccharomyces pombe:	
	1. SpCdc2:	"Cell-division-cycle" gene product
	Histoplasma capsulatum:	
*	1. HcCdc2:	Cdc2 homolog from dimorphic fungus
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. Pcdc2:	Flowering plant Cdc2 homolog othat complements yeast mutants
*	2. MsCdc2B:	Alfalfa Cdc2 cognate gene products that complements G1/S transition
	3. OsC2R:	More distantly related Cdc2 homolog from rice
	CMGC-II. Erk(MAP kinase) family	
	vertebrate:	
	1. Erk1:	"Extracellular signal-regulated kinase", type 1 (p44 MAP kinase)
	2. Erk2:	"Extracellular signal-regulated kinase", type 2 (p42 MAP kinase)
	3. Erk3:	Somewhat distant relative of the Erk/MAP kinases
*	4. p63MAPK:	Another more distant relative of the Erk/MAP kinases
*	5. SAPK-α:	"Stress-activated protein kinase, type alpha" (JNK2)
*	6. SAPK-β:	"Stress-activated protein kinase, type beta"
*	7. SAPK-γ/Jnk1:	"Stress-activated protein kinase, type gamma" or "Jun N-terminal Kinase"
*	8. p38:	HOG1-related protein (MPK2)
	Drosophila melanogaster:	
	1. DmErkA:	Homolog of Erk/MAP kinases; product of *rolled* gene
	Caenorhabditis elegans:	
*	1. Sur1:	Erk/MAP kinase homolog
	Saccharomyces cerrevisiae:	
	1. Kss1:	Suppressor of *sst2* mutant, overcomes growth arrest
	2. Fus3:	Product of gene required for growth and mating
	3. Slt2:	Product of gene complementing *lyt2* mutants (MPK1)
*	4. Hog1:	Product of gene required for osmoregulation
	Schizosaccharomyces pombe:	
	1. Spk1:	Product of gene that confers drug resistance to staurosporine, a PK inhibitor
	Phylum Deuteromycota (Kingdom Fungi):	
	1. CaErk1:	Protein that interferes with mating factor-induced cell cycle arrest
	Trypanosoma brucei (Phylum Zoomastigina, Kingdom Protoctista):	
*	1. KFR1:	"KSS1- and FUS3-related" gene product
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. PErk:	Flowering plant Erk/MAP kinase homologs (7 distinct homologs identified in Arabidopsis)
	CMGC-III. Glycogen synthase kinase 3 (GSK3) family	
	vertebrate:	
	1. GSK3α:	Glycogen synthase kinase 3, α-form
	2. GSK3β:	Glycogen synthase kinase 3, β-form
	Drosophila melanogaster:	
	1. Sgg:	Product of *shaggy/zeste-white 3* gene
	Saccharomyces cerevisiae:	
	1. Mck1:	"Meiosis and centromere regulatory kinase"
*	2. ScGSK3	Protein closely related to MCK1
*	3. Mds1:	Dosage suppressor of mck1 mutant
	Dictyostelium discoideum:	
*	1. DdGSK3:	Glycogen synthase kinase 3 homolog
	Phylum Angiospermophyta (Kingdom Plantae):	
*	1. ASK-α:	"Arabidopsis shaggy-related protein kinase", type alpha
*	2. ASK-γ:	"Arabidopsis shaggy-related protein kinase", type gamma

Table 1. *(continued).*

	vertebrate:	
	1. CK2α:	Casein kinase II, alpha subunit
	1. CK2α':	Casein kinase II, alpha-prime subunit
	Drosophila melanogaster:	
	1. DmCK2:	Casein kinase II homolog
	Caenorhabditis elegans:	
	1. CeCK2:	Casein kinase II homolog
	Theileria parva (a protozoan parasite):	
	1. TpCK2:	Casein kinase II α-subunit homolog
	Dictyostelium discoideum:	
	1. DdCK2:	Casein kinase II, α-subunit
	Saccharomyces cerevisiae:	
	1. ScCK2α:	Casein kinase II, alpha subunit
	2. ScCK2α':	Casein kinase II, alpha-prime subunit
	Schizosaccharomyces pombe:	
*	1. SpCka1:	Casein kinase II, α-subunit homolog (Orb5)
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. ZmCK2:	Flowering plant casein kinase II, α-subunit homolog
	CMGC-IV. Clk family	
	vertebrate:	
	1. Clk:	"Cdc-like kinase"
*	2. Srpk1:	Kinase that regulates intracellular localization of splicing factors
	3. PskG1:	Putative protein kinase
	4. PskH2:	Putative protein kinase
	Drosophila melanogaster:	
*	1. Doa:	Kinase encoded by "Darkener of Apricot" locus
	Saccharomyces cerevisiae:	
	1. Yak1:	Suppressor of RAS mutant
	2. Kns1:	Nonessential protein kinase homolog
	Schizosaccharomyces pombe:	
	1. Dsk1:	Dis1-suppressing protein kinase implicated in mitotic control
*	2. Prp4:	Pre-mRNA processing gene product; lacks subdomains X-XI
	Other CMGC Group kinases	
	vertebrate:	
	1. Mak:	"Male germ cell-associated kinase"
	2. Ched:	"Cholinesterase-related cell division controller"
	3. PITSLRE:	Galactosyltransferase-associated kinase
	4. KKIALRE:	Cdc2-related protein
*	5. PITALRE:	Cdc2-related kinase
*	6. PISSLRE:	Cdc2-related kinase
	Saccharomyces cerevisiae:	
	1. Sme1:	Product of gene essential for start of meiosis
	2. Sgv1:	Kinase required for G-protein-mediated adaptive response to pheromone
	3. Ctk1:	Product of gene required for normal growth
	Phylum Angiospermophyta (Kingdom Plantae):	
*	1. Mhk:	Arabidopsis thaliana "Mak homologous kinase"
	Conventional Protein-Tyrosine Kinase Group (I-X: Non-membrane-spanning; XI-XXIII: Membrane-spanning)	
	PTK-I. Src family	
	vertebrate:	
	1. Src:	Cellular homolog of Rous sarcoma virus oncoprotein
	2. Yes:	Cellular homolog of Yamaguchi 73 sarcoma virus oncoprotein
	3. Yrk:	Yes-related kinase
	4. Fyn:	Protein related to Fgr and Yes
	5. Fgr:	Cellular homolog of Gardner-Rasheed sarcoma virus oncoprotein
	6. Lyn:	Protein related to Fgr and Yes
	7. Hck:	Hematopoietic cell protein-tyrosine kinase
	8. Lck:	Lymphoid T-cell protein-tyrosine kinase
	9. Blk:	Lymphoid B-cell protein-tyrosine kinase
*	10. Frk:	Fyn-related kinase
*	11. Rak:	STK-related kinase
*	12. Fyk:	"Fyn and Yes-related kinase" from electric ray
	Drosophila melanogaster:	
	1. DmSrc:	Src homolog, polytene locus 64B
	Dugesiai (Girardia) tigrina (Phylum Platyhelminthes):	
*	1. DtSpk-1:	"Src-like planarian kinase"
	Hydra vulgaris (Phylum Cnidaria):	
	1. Stk:	Src-related protein
	Spongilla lacustris (Phylum Porifera):	
	1. Srk1-4:	Four distinct Src-related kinases
	PTK-II. Brk family	
	vertebrate:	
*	1. Brk:	Protein-tyrosine kinase expressed in human breast tumors

Table 1. *(continued)*.

	PTK-III. Tec family	
	vertebrate:	
	1. Tec:	"Tyrosine kinase expressed in hepatocellular carcinoma"
	2. Emt:	"Expressed mainly in T-cells" kinase (Itk, Tsk)
	3. Btk:	"Bruton's agammaglobulinaemia tyrosine kinase" (Emb)
*	4. Txk:	Tec-related protein-tyrosine kinase
	Drosophila melanogaster:	
	1. DmTec:	Tec homolog, polytene locus 28C
	PTK-IV. Csk family	
	vertebrate:	
	1. Csk:	"C terminal Src Kinase"; negative regulator of Src
*	2. MatK:	"Megakaryocyte-associated Tyr-kinase" (Hyl, Lsk, Ctk, Ntk)
	PTK-V. Fes(Fps) family	
	vertebrate:	
	1. Fes/Fps:	Cellular homolog of feline and avian sarcoma viruses
	2. Fer:	"Fes/Fps-related" kinase
	Drosophila melanogaster:	
	1. DmFer:	Fer-related protein
	PTK-VI. Abl family	
	vertebrate:	
	1. Abl:	Cellular homolog of Abelson murine leukemia virus
	2. Arg:	"Abl-related gene" product
	Drosophila melanogaster:	
	1. DmAbl:	Abl-related protein
	Caenorhabditis elegans:	
	1. CeAbl:	Nematode Abl-related protein
	PTK-VII. Syk/Zap70 family	
	vertebrate:	
	1. Syk:	"Spleen tyrosine kinase"
	2. Zap70:	T-cell receptor "zeta chain-associated protein of 70 kDa"
	Hydra vulgaris (Phylum Cnidaria):	
*	1. Htk16:	Syk/Zap70-related
	PTK-VIII. Jak family	
	vertebrate:	
	1. Tyk2:	Transducer of interferon α/β signals
	2. Jak1:	"Janus kinase", type 1
	3. Jak2:	"Janus kinase", type 2
*	4. Jak3:	"Janus kinase", type 3
	Drosophila melanogaster:	
*	1. Hop:	Product of hopscotch gene required for establishing segmental body plan
	PTK-IX. Ack	
	vertebrate:	
*	1. Ack:	"CDC42Hs-associated kinase"
	PTK-X. Fak	
	vertebrate:	
	1. Fak:	"Focal adhesion kinase"
	PTK-XI. Epidermal growth factor receptor family	
	vertebrate:	
	1. EGFR:	Epidermal growth factor receptor
	2. ErbB2:	Cell homolog of oncogene activated in ENU-induced rat neuroblastoma (Neu, HER2)
	3. ErbB3:	Receptor tyrosine kinase related to EGFR (HER3)
	4. ErbB4:	Receptor tyrosine kinase related to EGFR (Tyro2)
	Drosophila melanogaster:	
	1. DER:	Homolog of EGF receptor
	Caenorhabditis elegans:	
	1. LET-23:	Product of gene required for normal vulval development
	Schistosoma mansoni (Phylum Platyhelminthes):	
	1. SER:	EGF receptor homolog
	PTK-XII. Eph/Elk/Eck receptor family	
	vertebrate:	
	1. Eph:	Kinase detected in "erythropoeitin-producing hepatoma"
	2. Eck:	"Epithelial cell linase"
	3. Eek:	Eph/Elk-related protein-tyrosine kinase
	4. Hek:	Eph/Elk related protein-tyrosine kinase (Cek4)
	5. Sek:	"Segmentally-expressed kinase"
	6. Elk:	"Eph-like kinase" detected in brain
*	7. Hek2:	"Human embryo kinase" type 2 (Cek10)
*	8. Htk:	"Hepatoma transmembrane kinase"
	9. Cek5/Nuk:	"Chicken embryo kinase 5"/"Neural kinase"
*	10. Ehk1:	"Eph homology kinase-1" (Cek7)
*	11. Ehk2:	"Eph homology kinase-2"
*	12. Myk1:	"Mammary-derived tyrosine kinase, type 1"

Table 1. *(continued)*.

*	13. Myk2:	"Mammary-derived tyrosine kinase, type 2"
*	14. Cek9:	"Chicken embryo kinase 9"
*	15. Pag:	"Pagliaccio" Xenopus protein expression in neural crest and neural tissues
*	16. Rtk1:	Zebrafish Eph/Elk-related protein-tyrosine kinase
*	17. Rtk2:	Zebrafish Eph/Elk-related protein-tyrosine kinase
*	18. Rtk3:	Zebrafish Eph/Elk-related protein-tyrosine kinase
	PTK-XIII. Axl family	
	vertebrate:	
	1. Axl:	"Anexelekto" (Gr. "uncontrolled") tyrosine kinase (UFO, Ark)
	2. Eyk:	Cellular homolog of RPL30 avian oncoprotein (c-Ryk)
*	3. Brt/Sky/Tif/Rse:	"Brain tyrosine kinase"/"Sea related protein tyrosine kinase"/"Tyrosine kinase with Ig-like and FN-III-like domains"/"Receptor sectaris" (Tyro3)
	PTK-XIV. Tie/Tek family	
	vertebrate:	
	1. Tie:	"Tyrosine kinase with Ig and EGF homology"
	2. Tek:	"Tunica interna endothelial cell kinase" (TIE2)
	PTK-XV. Platelet-derived growth factor receptor family	
	A. Subfamily witih 5 Ig-like extracellular domains	
	vertebrate:	
	1. PDGFRα:	Platelet-derived growth factor receptor, type alpha
	2. PDGFRβ:	Platelet-derived growth factor receptor, type beta
	3. CSF1R:	Colony-stimulating factor-1 receptor (c-Fms)
	4. Kit:	Steel growth factor receptor
	5. Flk2:	"Fetal liver kinase-2" (Flt3)
	B. Subfamily with 7 Ig-like extracellular domains	
	vertebrate:	
	1. Flt1:	"Fms-like tyrosine kinase", type 1
	2. Flt4:	"Fms-like tyrosine kinase", type 4
	3. Flk1:	"Fetal liver kinase-1" (KDR)
	PTK-XVI. Fibroblast growth factor receptor family	
	vertebrate:	
	1. FGFR1:	Fibroblast growth factor receptor, type 1 (Flg, Cek1)
	2. FGFR2:	Fibroblast growth factor receptor, type 2 (Bek, K-SAM, Cek3)
	3. FGFR3:	Fibroblast growth factor receptor, type 3
	4. FGFR4:	Fibroblast growth factor receptor, type 4
	Drosophila melanogaster:	
	1. DmFGFR1:	Fibroblast growth factor receptor homolog, type 1
*	2. DmFGFR2:	Fibroblast growth factor receptor homolog, type 2
	PTK-XVII. Insulin receptor family	
	vertebrate:	
	1. InsR:	Insulin receptor
	2. IGF1R:	Insulin-like growth factor receptor
	3. IRR:	Insulin receptor-related protein
	Drosophila melanogaster:	
	1. DmInsR:	Homolog of insulin receptor
	PTK-XVIII. Ltk/Alk family	
	vertebrate:	
	1. Ltk:	"Leukocyte tyrosine kinase
*	2. Alk:	"Anaplastic lymphoma kinase
	PTK-XIX. Ros/Sev family	
	vertebrate:	
	1. Ros:	Cellular homolog of UR2 avian sarcoma virus oncoprotein
	Drosophila melanogaster:	
	1. Sev:	Product of *sevenless* gene required for R7 photoreceptor cell development
	PTK-XX. Trk/Ror family	
	vertebrate:	
	1. Trk:	High molecular weight nerve growth factor receptor
	2. TrkB:	Receptor for nrain-derived neurotrophic factor and neurotrophin-4/5
	3. TrkC:	Trk-related protein; receptor for neurotrophin-3
	4. Ror1:	"Ror" putative receptor, type 1
	5. Ror2:	"Ror" putative receptor, type 2
	6. TcRTK:	Trk-related receptor (electric ray)
	Drosophila melanogaster:	
*	1. Dror:	Putative neurotrophic receptor
	PTK-XXI. Ddr/Tkt family	
*	1. Ddr:	"Discoidin Domain Receptor" (TrkE, CAK, NEP, Ptk3)
*	2. Tkt:	"Tyrosine Kinase Related to Trk" (Tyro 10)

Table 1. *(continued).*

PTK-XXII. Hepatocyte growth factor receptor family		
vertebrate:		
	1. HGFR:	Hepatocyte growth factor receptor (MET)
	2. Sea:	Cellular homolog of S13 avian erythroleukemia virus oncoprotein
	3. Ron:	"Recepteur d'Origine Nantaise"
*	4. Stk:	"Stem cell-derived tyrosine kinase"
PTK-XXIII. Nematode Kin15/16 family		
Caenorhabditis elegans:		
	1. CeKin15:	PTK expressed during hypodermal development
	2. CeKin16:	PTK expressed during hypodermal development
Other membrane-spanning protein-tyrosine kinases (each with no close relatives)		
vertebrate:		
	1. Ret:	Normal homolog of oncoprotein activated by recombination
	2. Klg:	"Kinase-like gene" product
*	3. Nyk/Ryk:	"Novel tyrosine kinase-related protein" (VIK, Mrk, Nbtk1)
Drosophila melanogaster:		
	1. Torso:	Product of *torso* gene required for embryonic anterior/posterior determination
	2. DmTrk:	Distant relative of the mammalian trk gene
Marine sponge (Geodia cydonium):		
*	1. GCTK:	Putative receptor PTK
Other protein kinase families (not falling into m ajor groups)		
O-I. Polo family		
vertebrate:		
	1. Plk:	"Polo-like kinase"
	2. Snk:	"Serum-inducible kinase"
*	3. Sak:	Polo-related kinase isolated in screen for genes regulating sialylation
Drosophila melanogaster:		
	1. Polo:	Protein kinase homolog required for mitosis
Saccharomyces cerevisiae:		
	1. Cdc5:	Product of gene required for cell cycle progression
O-II. MEK/STE7 family		
vertebrate:		
	1. MEK1:	"MAP ERK Kinase", type 1
	2. MEK2:	"MAP ERK Kinase", type 2
Drosophila melanogaster:		
	1. Dsor1:	
Saccharomyces cerevisiae:		
	1. Ste7:	Kinase required for haploid-specific gene expression
	2. Pbs2:	Kinase required for antibiotic drug resistance
	3. Mkk1:	"MAP Kinase Kinase", type 1 (suppresses lysis defect of pkc1 mutant)
	4. Mkk2:	"MAP Kinase Kinase", type 2 (suppresses lysis defect of pkc1 mutant)
Schizosaccharomyces pombe:		
	1. Byr1:	Kinase that suppresses ras1-mutant sporulation defect
	2. Wis1:	Suppressor of cdc phenotype in triple mutant *cdc25/wee1/win1* strains
O-III. MEKK/Ste11 family		
vertebrate:		
*	1. MEKK:	"MEK Kinase"
Saccharomyces cerevisiae:		
	1. Ste11:	Protein required for cell-type-specific transcription
	2. Bck1:	"Bypass of C kinase" kinase
Schizosaccharomyces pombe:		
	1. Byr2:	Product of gene required for pheromone signal transduction
Phylum Angiospermophyta (Kingdom Plantae):		
*	1. NPK1:	Flowering plant (tobacco) homolog of Bck1
O-IV. Pak/Ste20 family		
vertebrate:		
*	1. Pak:	"p21-(Cdc42/Rac) activated kinase"
Saccharomyces cerevisiae:		
	1. Ste20:	Product of gene required for pheromone response
O-V. NimA family		
vertebrate:		
	1. Nek1:	NimA-related kinase
*	2. Nek2:	NimA-related kinase (Nlk1)
*	3. Nek3:	NimA-related kinase
*	4. Nrk2:	NimA-related kinase
*	5. Stk1:	NimA-related kinase
Aspergillus nidulans:		
	1. NIMA:	Cell cycle control protein kinase
Drosophila melanogaster:		
	1. Fused:	Product of gene required for segment polarity

Table 1. *(continued).*

	Trypanosoma brucei (Phylum Zoomastigina, Kingdom Protoctista):	
	1. NrkA:	Trypanosome protein kinase related to NimA
	Saccharomyces cerevisaie:	
	1. Kin3:	Putative protein kinase
	O-VI. wee1/mik1 family	
	vertebrate:	
	1. Wee1Hu:	Gene product able to complement S. pombe wee1 mutant
	Saccharomyces cerevisiae:	
*	1. Swe1:	Wee1 homolog from budding yeast
	Schizosaccharomyces pombe:	
	1. SpWee1:	"Wee" size at division kinase; Cdc2 negative regulator
	2. Mik1:	"Mitosis inhibitory kinase", negative regulator of Cdc2
	O-VII. Family of kinases involved in translational control	
	vertebrate:	
	1. HRI:	"Heme-regulated eukaryotic initiation factor 2α kinase"
	2. PKR:	"Double-stranded RNA-dependent kinase" (Tik)
	Saccharomyces cerevisiae:	
	1. Gcn2:	Protein required for translational derepression
	O-VIII. Raf family	
	vertebrate:	
	1. Raf-1:	Cellular homolog of retroviral oncogene product
	2. A-Raf:	Oncogenic protein closely related to c-Raf
	3. B-Raf:	Oncogenic protein closely related to c-Raf
	Drosophila melanogaster:	
	1. DmRaf:	Raf homolog
	Caenorhabditis elegans:	
	1. CeRaf:	Raf homolog; product of *lin-45* gene required for vulval differentiation
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. Ctr1:	Negative regulator of ethylene response pathway
	O-IX. Activin/TGFβ receptor family	
	A. Subfamily of type I receptors	
	vertebrate:	
	1. ActR-I:	Type I receptor for activin and TGF-β (Tsk7L, SKR1, ALK-2)
*	2. TSR-1:	Type I receptor for activin and TGFG-β (ALK-1)
*	3. TGFβRI:	Type I receptor TGF- (ALK-5)
*	4. ActR-IB:	Type I receptor for activin (ALK-4)
*	5. BRK-1:	Type I receptor for BMP-2 and BMP-4 (ALK-3)
*	6. ALK-6:	"Activin receptor-like kinase", type 6
	Drosophila melanogaster:	
*	1. DmAtr-I:	Type I activin receptor homolog
*	2. DmSax:	Product of *saxophone* gene
	B. Subfamily of type II receptors	
	vertebrate:	
	1. ActRII:	Type II receptor for activin
	2. ActRIIB:	Type II receptor for activin
	3. TGFβRII:	Type II receptor TGF-β
*	4. C14:	Putative receptor kinase expressed in gonads
	Drosophila melanogaster:	
*	1. DmAtr-II:	Type II activin receptor homolog
	Caenorhabditis elegans:	
*	1. DAF-4:	Larva development regulatory protein; BMP receptor
	C. Others	
	Caenorhabditis elegans:	
	1. DAF-1:	Product of gene required for vulval development
	O-X. Flowering plant putative receptor kinase family	
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. ZmPK1:	Putative receptor protein-serine kinase (maize)
	2. Srk:	"S receptor kinase"; three distinct alleles: 2, 6, and 910 (Brassica)
	3. Tmk1:	Putative "Transmembrane receptor kinase" (Arabidopsis)
	4. Apk1:	Kinase that phosphorylates Tyr, Ser, and Thr (Arabidopsis)
*	5. Nak:	"Novel Arabidopsis Kinase" (Arabidopsis)
	6. Pro25:	Putative kinase selected for specificity to thylakoid membrane protein (Arabidopsis)
*	7. Pto:	Product of genen conferring pathogen resistance (tomato)
*	8. Tmkl1:	Transmembrane protein with unusual kinase-like domain (Arabidopsis)
*	9. Prk1:	Pollen-expressed receptor-like putative kinase (Petunia)
	O-XI. Family of "mixed-lineage" kinases with leucine zipper domain	
	vertebrate:	
*	1. Mlk1:	"Mixed lineage kinase", type 1
*	2. Mlk2:	"Mixed lineage kinase", type 2
*	3. Mlk3:	"Mixed lineage kinase", type 3 (PTK1, SPRK)

Table 1. *(continued).*

	O-XII. Casein kinase I family	
	vertebrate:	
	1. CK1α:	Casein kinase I, type alpha
	2. CK1β:	Casein kinase I, type beta
	3. CK1γ:	Casein kinase I, type gamma
	4. CK1δ:	Casein kinase I, type delta
	Saccharomyces cerevisiae:	
	1. Yck1:	Budding yeast casein kinase I homolog, type 1
	2. Yck2:	Budding yeast casein kinase I homolog, type 2
	3. Hrr25:	Kinase required for DNA repair
	Schizosaccharomyces pombe:	
*	1. Hhp1:	Fission yeast casein kinase I homolog, type 1
*	2. Hhp2:	Fission yeast casein kinase I homolog, type 2
	O-XIII. PKN family of prokaryotic protein kinases	
	Myxococcus xanthus (Phylum Myxobacteria: Kingdom Prokaryotae):	
	1. Pkn1:	Protein kinase homologous to eukaryotic kinases
	2. Pkn2:	Protein kinase required for maintenance of stationary phase cells and development
	Other protein kinase family members (each with no known close relatives)	
	vertebrate:	
	1. Mos:	Cellular homolog of retroviral oncogene product
	2. Pim1:	Proto-oncogene activated by murine leukemia virus
	3. Cot:	Product of oncogene expressed in human thyroid carcinoma
	4. Esk:	"Embryonal carcinoma STY kinase"; dual specificity (PIT)
*	5. GC kinase:	Kinase expressed in germinal center B cells
*	6. Slk:	STE20-related kinase
*	7. LIMK:	"LIM motif-containing kinase"
*	8. Tsk1:	"Testis-specific kinase"
	Drosophila melanogaster:	
	1. NinaC:	Product of gene essential for photoreceptor function
	2. Pelle:	Product of gene required for dorsalventral polarity
*	3. Nemo:	Product of gene required for rotation of photoreceptor clusters
	Dictyostelium discoideum:	
	1. Spl A:	Spore lysis A protein kinase
	2. Dpyk2:	Developmentally-reguated tyrosine kinase, type 2
	Ceratodon purpureus: (a moss)	
	1. PhyCer:	Putative protein-tyrosine kinase encoded by a phytochrome gene
	Saccharomyces cerevisiae:	
	1. Cdc7:	"Cell-division-cycle" control gene product
	2. CDC15:	"Cell-division-cycle" control gene product
	3. Vps15:	Product of gene essential for sorting to lysosome-like vacuole
	4. Npr1:	Product of gene required for activity of ammonia-sensitive amino acid permeases
	5. Elm1:	Product of gene required for yeast-like cell morphology
	6. Ire1:	Required for Myo-inositol synthesis and signaling from ER to the nucleus
	7. Ykl516:	Putative protein kinase gene on chromosome XI
*	8. Ipl1:	Product of gene required for chromosome segregation
	Schizosaccharomyces pombe:	
	1. Ran1:	Product of gene required for normal meiotic function
	2. Chk1:	"Checkpoint Kinase" that links rad pathway to Cdc2
*	3. Csk1:	"Cyclin Suppressing Kinase"
*	4. RPK1:	"Regulatory cell proliferation kinase"
	Entamoeba histolytica (Phylum Rhizopoda, Kingdom Protoctista):	
	1. Ehmfk1:	Distant relative of Mos
	Phylum Angiospermophyta (Kingdom Plantae):	
	1. GmPK6:	Protein kinase homolog (soybean)
*	2. Tsl:	Product of *Tousled* gene required for normal leaf/flower development (Arabidopsis)
	Yersinia psuedotuberculosis (Phylum Omnibacteria, Kingdom Prokaryotae):	
	1. YpkA:	Enterobacterial protein kinase essential for virulence

known primary structures. The kinase domains are further divided into 12 smaller subdomains (indicated by Roman numerals), defined as regions never interrupted by large amino acid insertions and containing characteristic patterns of conserved residues (consensus line in Fig. 1).

Twelve kinase domain residues are recognized as being invariant or nearly invariant throughout the superfamily (conserved in over 95% of 370 sequences), and hence strongly implicated as playing essential roles in enzyme function. Using the type α cAMP-dependent protein kinase catalytic subunit (PKA-Cα) as a reference point, these are equivalent to Gly50 and Gly52 in subdomain I, Lys72 in subdomain II, Glu91 in subdomain III, Asp166 and Asn171 in subdomain VIB, Asp184 and Gly186 in subdomain VII, Glu208 in subdomain VIII, Asp220 and Gly225 in subdomain IX, and Arg280 in subdomain XI.

The patterns of amino acid residues found within subdomains VIB, VIII, and IX have been particularly well-conserved among the individual members of the dif-

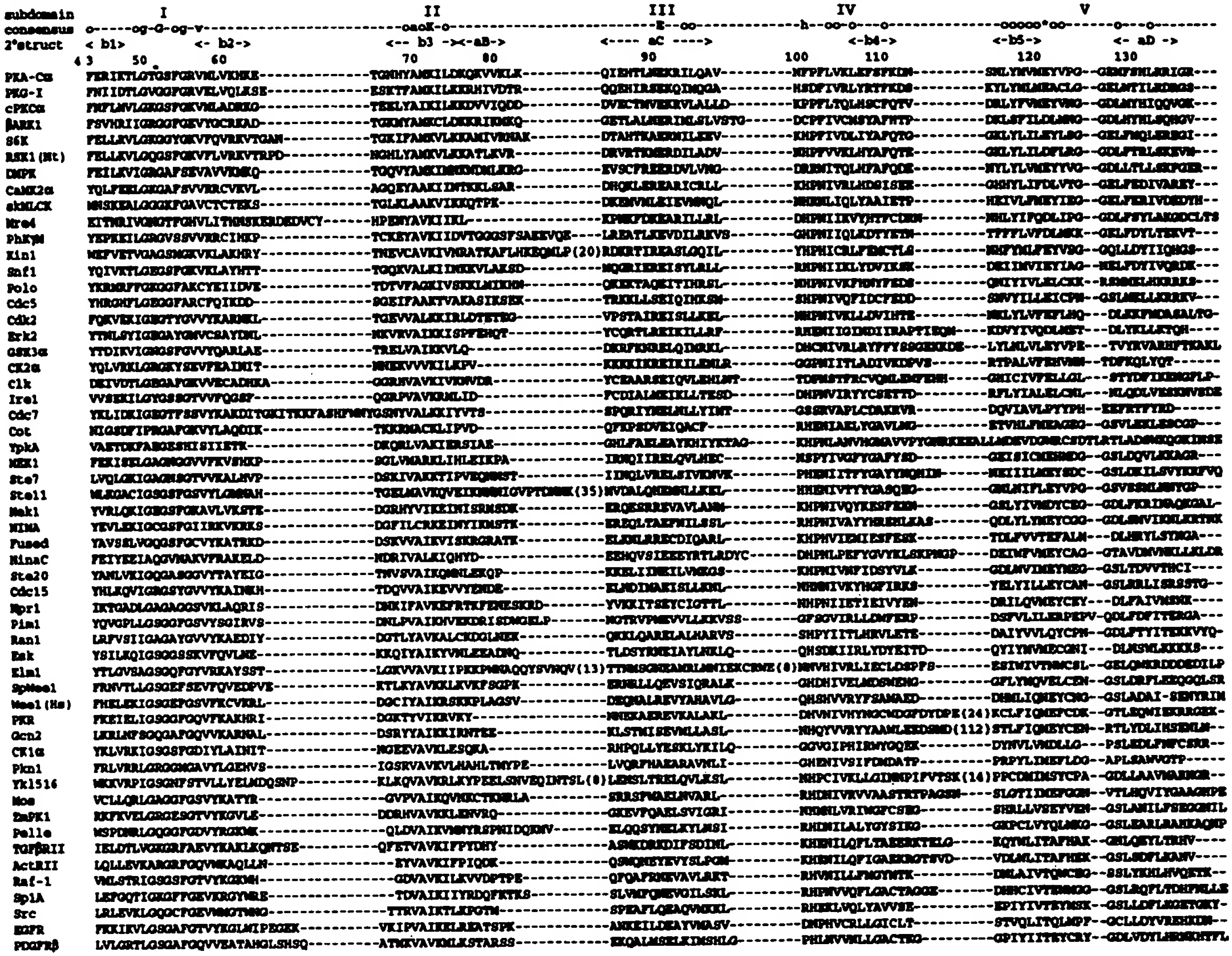

subdomain
consensus
2°struct
I
II
III
IV
V
PKA-Cα
PKG-I
cPKCα
βARK1
S6K
RSK1(Nt)
DMPK
CaMK2α
skMLCK
Mre4
PhKγM
Kin1
Snf1
Polo
Cdc5
Cdk2
Erk2
GSK3α
CK2α
Clk
Ire1
Cdc7
Cot
YpkA
MEK1
Ste7
Ste11
Mak1
NIMA
Fused
NinaC
Ste20
Cdc15
Mpr1
Pim1
Ran1
Esk
Elm1
SpWee1
Wee1(Hs)
PKR
Gcn2
CK1α
Pkn1
Yk1516
Mos
ZmPK1
Pelle
TGFβRII
ActRII
Raf-1
SplA
Src
EGFR
PDGFRβ

Figure 1. Multiple alignments of 60 kinase domains representative of members of the eukaryotic protein kinase superfamily. The abbreviated names used are as defined in Table 1. The single letter amino acid code is used and gaps are indicated by dashes. The entire sequences for the larger inserts are not shown, but excluded residues are indicated as numbers in brackets. Twelve distinct subdomains are indicated by Roman numerals. The consensus line is given according to the following code: uppercase letters, invariant residues, lowercase residues nearly invariant residues; o, positions conserving nonpolar residues; *, positions conserving polar residues; +, positions conserving small residues with near neutral polarity. Residues corresponding to the numbered β-strands (b) and α-helices (a) in PKA-Cα are indicated in the 2• structure line.

ferent protein kinase families and these motifs have been targeted most frequently in PCR-based homology cloning strategies aimed at identifying new family members.

Relationship between conserved subdomains, higher order structure, and catalytic mechanism

The homologous nature of the kinase domains implies that they all fold into topologically similar 3-dimensional core structures and impart phosphotransfer according to a common mechanism. The larger inserts found within some kinase domains are likely to represent surface elements that do not disrupt the basic core structure. With the solution of the crystal structure of mouse PKA-Cα, in a binary complex with a pseudosubstrate peptide inhibitor (PKI 5-24; TTYADFIASGRTGRRNAIHD, the underlined Ala substituting for the Ser phosphoacceptor), the general topology of a protein kinase catalytic core structure was revealed for the first time (25, 26). Later, structures of ternary complexes of PKA-Cα, the pseudosubstrate inhibitor, and either MgATP or MnAMP-PNP (an MgATP analog) were solved (27, 28). As a consequence of these studies, precise functional roles for most of the highly conserved kinase domain residues have now been assigned.

The kinase domain of PKA-Cα folds into a two-lobed structure (**Fig. 2**). The smaller, NH_2- terminal lobe, which includes subdomains I-IV, is primarily involved in anchoring and orienting the nucleotide. This lobe has a predominantly antiparallel β-sheet structure that is unique among nucleotide binding proteins. The larger COOH-terminal lobe, which includes subdomains VIA-XI, is largely responsible for binding the peptide substrate and initiating phosphotransfer. It is predominantly α-helical in content. Subdomain V residues span

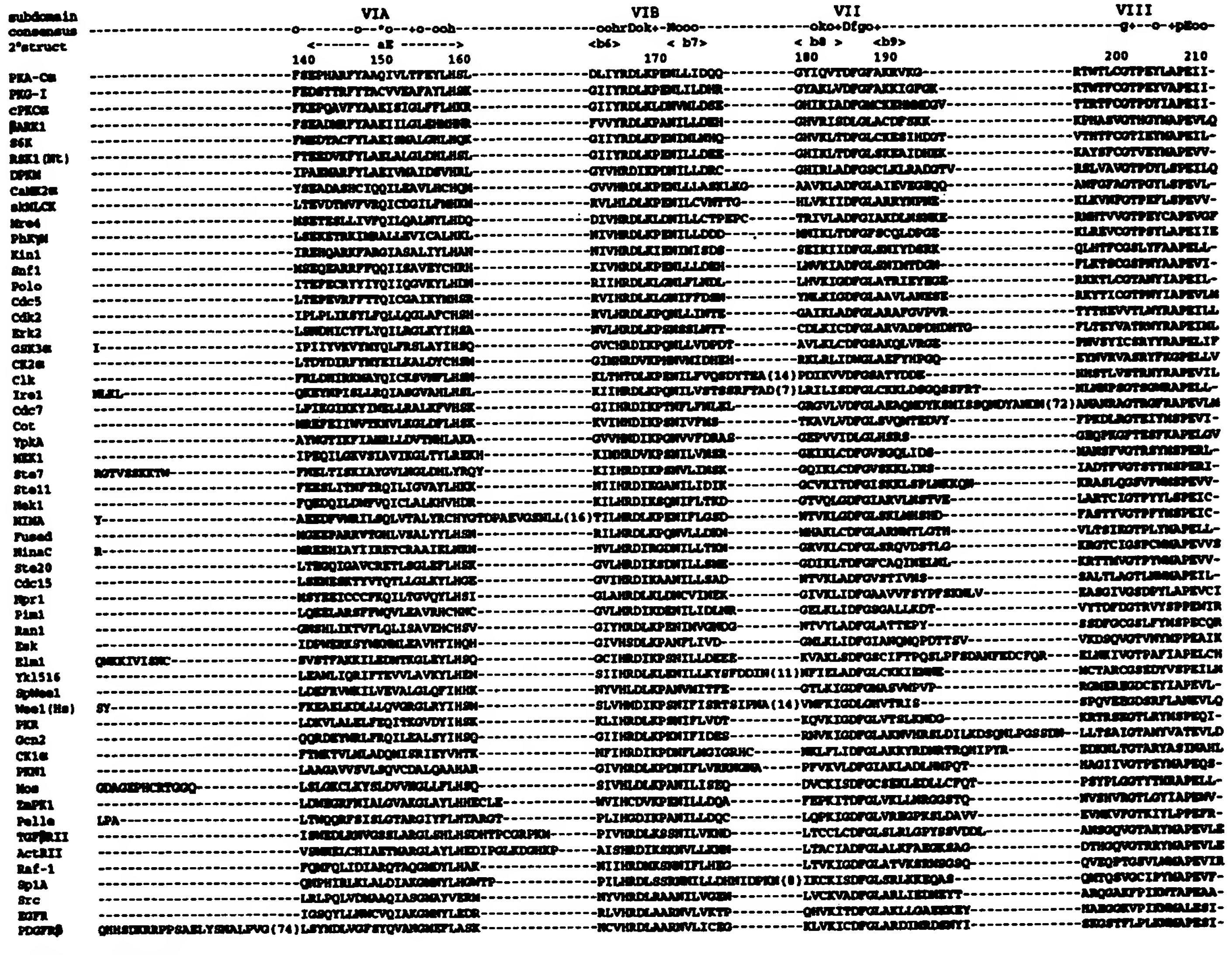


Figure 1 (contd.).

the two lobes. The deep cleft between the two lobes is recognized as the site of catalysis. The crystal structures of four additional eukaryotic protein kinase superfamily members—cyclin-dependent kinase 2 (Cdk2) (29), p42 MAP kinase (Erk2) (30), twitchin kinase (31), and casein kinase I (32)—have been reported more recently, and as expected, their kinase domains were found to fold into two-lobed structures topologically very similar to the catalytic core of PKA-Cα. Notable differences, however, were found in the regions corresponding to subdomain VIII in the Cdk2 and Erk2 structures, apparently reflecting the fact that these are structures of enzymes in an inactive state (see below). The twitchin structure is also of an inactive enzyme, but in this case it is inactive due to the presence of an autoinhibitory peptide sequence, which lies on the COOH-terminal side of the kinase domain and folds back into the active site cleft between the two lobes (31). This peptide apparently forces the two lobes to rotate almost 30° with respect to one another, and in this configuration inactive twitchin is more similar to the open configuration of PKA-Cα without PKI (33). In both twitchin and Cdk2 the α-helix C in subdomain III also adopts a different position to that of helix C in PKA-Cα. Unfortunately, no structure of a protein-tyrosine kinase catalytic domain was available at the time of writing (see "Note added in proof"), but the ease with which it has been possible to model the kinase domain of the EGF receptor protein-tyrosine kinase on to that of the PKA-Cα emphasizes that the structure of the protein-tyrosine kinases will be similar to that of the protein-serine kinases (34)

The conserved kinase subdomains correspond quite well to precise units of higher order structure. The functions of the individual subdomains will be discussed briefly later on a subdomain-by-subdomain basis, making reference to the crystal structure of PKA-Cα and

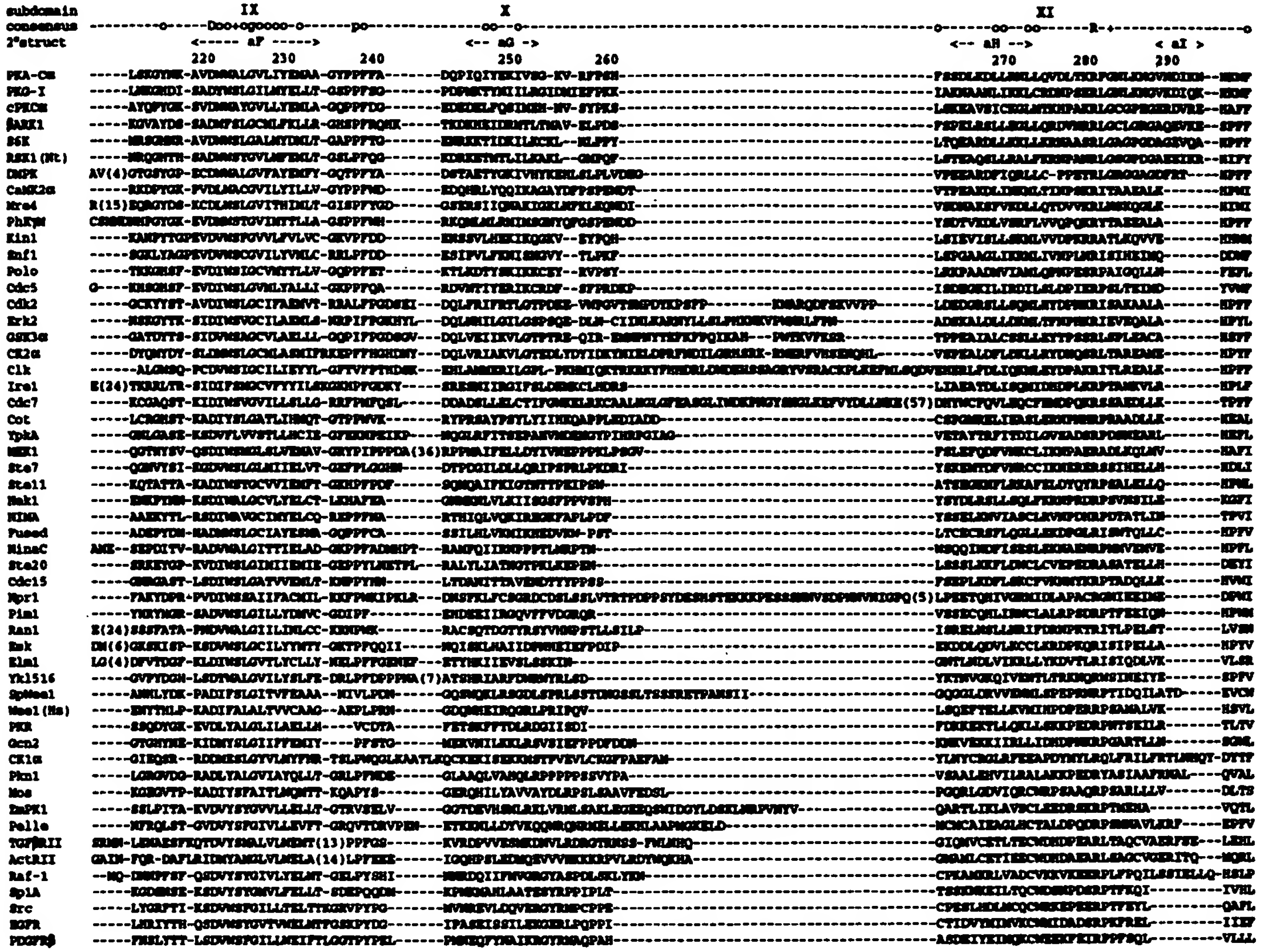


Figure 1 (contd.).

drawing attention to the proposed roles of the nearly invariant amino acid residues (25-27, 28) and other residues of interest. For more detailed information, the reader is referred to recent reviews on the structure of PKA-Cα (35-37) and to an excellent comparative review of the structures of PKA-Cα, Erk2, and Cdk2 (38).

Subdomain I, at the NH_2 terminus of the kinase domain, contains the consensus motif Gly-x-Gly-x-x-Gly-x-Val (starting with Gly50 in PKA-Cα). The kinase domain NH_2-terminal boundary occurs seven positions upstream of the first glycine in the consensus, where a hydrophobic residue is usually found. Subdomain I residues fold into a β-strand-turn-β-strand structure encompassing β-strands 1 and 2, and this structure acts as a flexible flap or clamp that covers and anchors the nontransferable phosphates of ATP. The backbone amides of Ser53, Phe54, and Gly55 form hydrogen bonds with ATP β- phosphate oxygens. Leu49 and Val57 contribute to a hydrophobic pocket that encloses the adenine ring of ATP.

Subdomain II contains the invariant Lys (Lys72 in PKA-Cα), which has long been recognized as being essential for maximal enzyme activity. This Lys lies within β-strand 3 of the small lobe, and helps anchor and orient ATP by interacting with the α- and β- phosphates. In addition, Lys72 forms a salt bridge with the carboxyl group of the nearly invariant Glu91 in subdomain III. Ala70 contributes to the hydrophobic adenine ring pocket. In PKA-Cα, β-strand 3 is followed immediately by α-helix B, which, judging from the sequence alignment, appears to be quite a variable structure among the protein kinases. Indeed, this α- helix is absent in the Cdk2 and Erk2 crystal structures.

Subdomain III represents the large α- helix C in the small lobe. The nearly invariant Glu residue (Glu91 in PKA-Cα) is centrally located in this helix and helps stabilize the interactions between Lys72 and the α- and β- phosphates of ATP. Subdomain IV corresponds to the hydrophobic β-strand 4 in the small lobe. This subdomain contains no invariant or nearly invariant residues

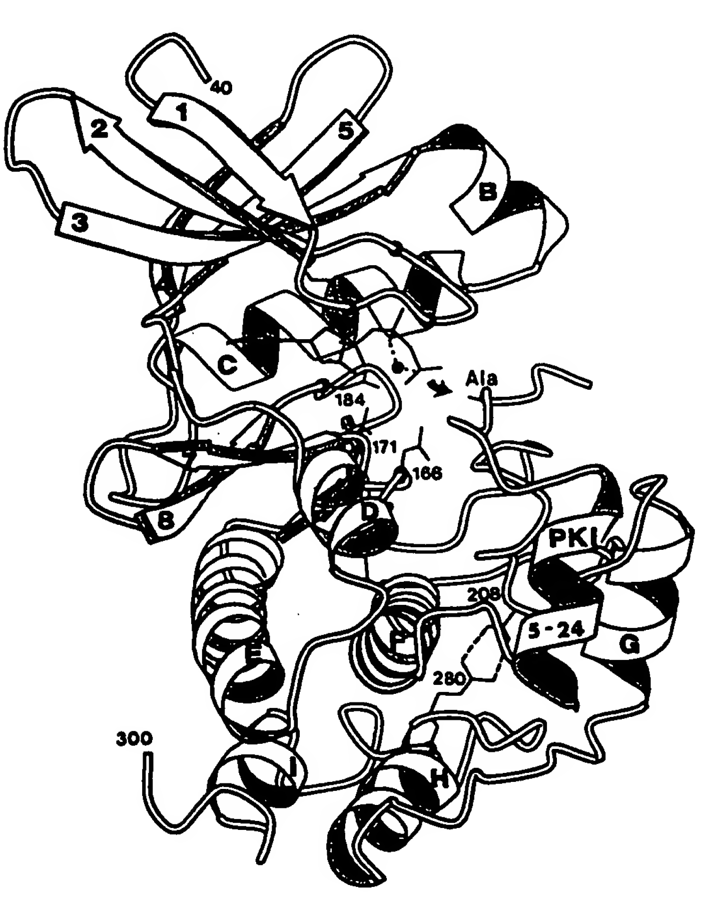


Figure 2. Ribbon diagram of the catalytic core of PKAα (residues 40–300) in a ternary complex with MgATP and pseudosubstrate peptide inhibitor (PKI -5–24). Invariant or nearly-invariant residues (Gly50, Gly52, Gly55, Lys72, Glu91, Asp166, Asn171, Asp184, Glu208, Asp220, and Arg280) are indicated by dots along the ribbon diagram. Side chains are shown for Lys72, Asp166, Asn171, Asp184, Glu208, and Arg280. β-strands and α-helices are indicated by flat arrow and helices, respectively, and are numbered according to Knighton et al. (26). The small arrow indicates the site of phosphotransfer with the Ala in PKI substituting for the phosphoacceptor Ser in the true substrate. (Reproduced, with permission, from Taylor et al. (36)).

and does not appear to be directly involved in catalysis or substrate recognition.

Subdomain V links the small and large lobes of the catalytic subunit and consists of the very hydrophobic β-strand 5 in the small lobe, the small α-helix D in the large lobe, and an extended chain that connects them. Three residues in the connecting chain of PKA-Cα, Glu121, Val123, and Glu127 help anchor ATP by forming hydrogen bonds with either the adenine or the ribose ring. Met120, Tyr122, and Val123 contribute to the hydrophobic pocket surrounding the adenine ring. Glu127 also participates in peptide binding by forming an ion pair with an Arg in the pseudosubstrate site of the PKA inhibitor peptide. This represents the first Arg in the PKA substrate recognition consensus Arg-Arg-x-Ser*-Hydrophobic.

Subdomain VIA folds into the large hydrophobic α-helix E that extends through the large lobe. None of the residues in helix E appear to interact directly with either MgATP or peptide substrate; hence this part of the molecule appears to act mainly as a support structure. Subdomain VIB folds into the small hydrophobic β-strands 6 and 7 with an intervening loop. Included here are two invariant residues (Asp166 and Asn171 in PKA-Cα) that lie within the consensus motif His-Arg-Asp-Leu-Lys-x-x-Asn (HRDLKxxN). The loop has been termed the catalytic loop because Asp166 within the loop has emerged as the likely candidate for the catalytic base, accepting the proton from the attacking substrate hydroxyl group during an in- line phosphotransfer mechanism. Lys168 in the loop (substituted by Arg in the conventional protein-tyrosine kinases) may help facilitate phosphotransfer by neutralizing the negative charge of the γ-phosphate during transfer. The side chain of Asn171 helps to stabilize the catalytic loop through hydrogen bonding to the backbone carbonyl of Asp166 and also acts to chelate the secondary Mg^{2+} ion that bridges the α- and γ-phosphates of the ATP. The carbonyl group of Glu170 forms a hydrogen bond with an ATP ribose hydroxyl group. Glu170 also participates in substrate binding by forming an ion pair with the second arginine of the peptide recognition consensus.

Subdomain VII folds into a β-strand-loop-b-strand structure, encompassing β-strands 8 and 9. The highly conserved DFG triplet, corresponding to Asp184-Phe185-Gly186 in PKA-Cα, lies in the loop that is stabilized by a hydrogen bond between Asp184 and Gly186. Asp184 chelates the primary activating Mg^{2+} ions that bridge the β- and γ-phosphates of the ATP, and thereby helps to orient the γphosphate for transfer. In Cdk2, β-strand 9 is replaced with a small α-helix designated αL12. However, it is unclear whether this helical character is maintained when Cdk2 is in its active conformation.

Subdomain VIII, which includes the highly conserved Ala-Pro-Glu ('APE') motif (residues 206–208 in PKA-Cα), folds into a tortuous chain that faces the cleft. Residues lying 7–10 positions immediately upstream of the APE motif are characteristically well-conserved among the members of different protein kinase families. The nearly invariant Glu corresponding to PKA-Cα Glu208 forms an ion pair with an invariant Arg (Arg280 in PKA-Cα) in subdomain XI, thereby helping to stabilize the large lobe.

Subdomain VIII appears to play a major role in recognition of peptide substrates. Several PKA-Cα subdomain VIII residues participate in binding the pseudosubstrate inhibitor peptide. Leu198, Cys199, Pro202, and Leu205 of PKA-Cα provide a hydrophobic pocket that accommodates the side chain of the hydrophobic residue at position +1 of the substrate consensus (Ile for the inhibitor peptide). Gly200 forms a hydrogen bond with the same Ile residue. Glu203 forms two ion pairs with the Arg in the high-affinity binding region of the inhibitor peptide.

Many protein kinases are known to be activated by phosphorylation of residues in subdomain VIII. In PKA-Cα, maximal kinase activity requires phosphorylation of Thr197, probably occurring through an intermolecular autophosphorylation mechanism (39). In the crystal structure, phosphate oxygens of phospho-Thr197 form hydrogen bonds with the charged side chains of Arg165, Lys189, and the hydroxyl group of Thr195, and thereby may act to stabilize the subdomain VIII loop in an active conformation permitting proper orientation of the substrate peptide. For members of the Erk (MAP) kinase family, phosphorylation of both a Thr and a Tyr residue in subdomain VIII (mediated by members of the MEK kinase family) is required for activation. In the crystal structure determined for Erk2, these residues (Thr183 and Tyr185) were not phosphorylated and thus the enzyme was in an inactive state (unlike the PKA-Cα structure). The unphosphorylated Tyr185 is buried in a hydrophobic pocket, and interactions with Tyr185 are apparently required to hold the enzyme in the inactive state. Mutation of Tyr185, however, does not activate the enzyme, and so phosphorylation of Tyr185 must also play a role in activation. Unphosphorylated Erk2 appears to be inactive because residues required for catalysis are not properly oriented, and because its conformation results in a partial steric block to substrate binding. During activation of Erk2, Tyr185 phosphorylation precedes Thr183 phosphorylation; therefore, binding of MEK to Erk2 may alter the conformation of the subdomain VIII loop, thereby exposing Tyr185 for phosphorylation by MEK. Interaction of phospho-Tyr185 with surface residues would then allow the subdomain VIII loop to adopt the active conformation (30). Subsequent phosphorylation of the exposed Thr183 may activate the enzyme fully by promoting correct alignment of the catalytic residues. From the crystal structure of Cdk2, likewise in an inactive unphosphorylated state, the subdomain VIII loop appears to be in a conformation that would inhibit enzyme activity by sterically blocking the presumed protein substrate binding cleft (29). Phosphorylation of Thr160 in the Cdk2 subdomain VIII, mediated by MO15 (CAK), presumably would act to remove this inhibition by stabilizing the loop in an active conformation similar to that found in PKA-Cα. Cyclin binding to the NH_2-terminal lobe is also needed to activate Cdk2, and this may cause rotation of the NH_2-terminal domain resulting in correct alignment of catalytic residues.

Subdomain IX corresponds to the large α- helix F of the large lobe. The nearly invariant Asp corresponding to PKA-Cα Asp220 lies in the NH_2-terminal region of this helix and acts to stabilize the catalytic loop by hydrogen bonding to the backbone amides of Arg165 and Tyr164 that precede the loop. Glu230 of PKA-Cα forms an ion pair with the second Arg of the peptide recognition consensus. PKA-Cα residues 235-239 are all involved in hydrophobic interactions with the inhibitor peptide.

Subdomain X is the most poorly conserved subdomain and its function is obscure. In the crystal structure of PKA-Cα, it corresponds to the small α-helix G that occupies the base of the large lobe. Members of the Cdk, Erk (MAP), GSK3, and Clk kinase families (the C-M-G-C group) all have rather large insertions between subdomains X and XI, whose functional significance is presently unclear. Subdomain XI extends to the COOH-terminal end of the kinase domain. The most notable feature here is the nearly invariant Arg corresponding to Arg280 in PKA-Cα, which lies between α-helices H and I. The COOH-terminal boundary of the kinase domain is still poorly defined. For many protein-serine kinases, the consensus motif His-x-Aromatic-Hydrophobic is found beginning 9-13 residues downstream of the invariant Arg. For protein-tyrosine kinases, a hydrophobic amino acid lying 10 positions downstream of the invariant Arg appears to define the COOH-terminal boundary.

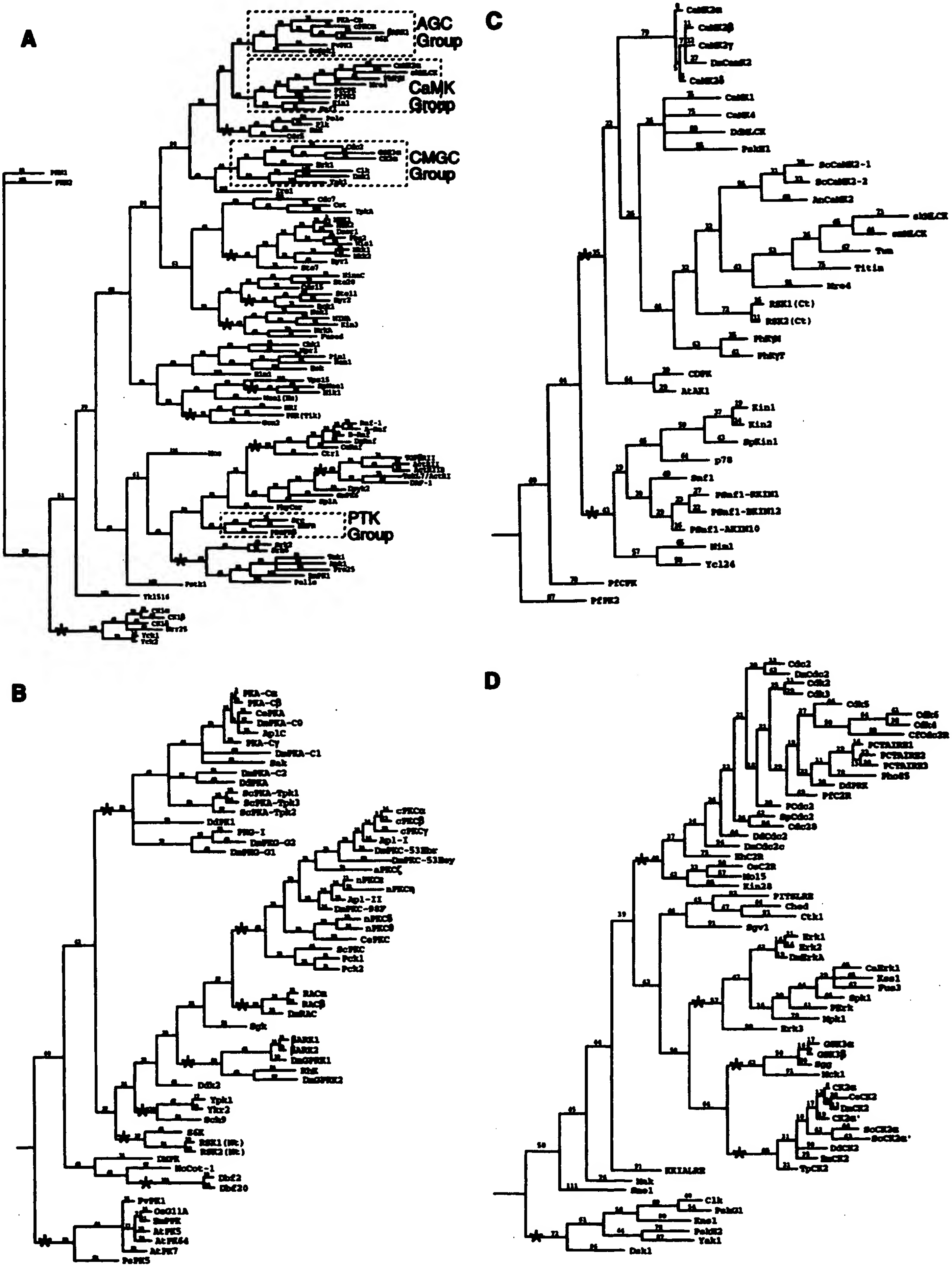
The amphipathic α-helix A of PKA-Ca (residues 15-35; not shown in Fig. 2), though lying outside of the conserved catalytic core on the NH_2-terminal side, appears to be an important feature found in many protein kinases (40). This helix spans the surface of both lobes of the core structure and complements and stabilizes the hydrophobic cleft between the two lobes. The A-helix motif appears to be present in many other protein kinases including members of the protein kinase C family and the Src family of protein-tyrosine kinases (40).

CLASSIFICATION OF EUKARYOTIC PROTEIN KINASES

To facilitate analysis and management of this large superfamily we have devised the classification scheme shown in Table 1, which subdivides the known members of the eukaryotic protein kinase superfamily into distinct families that share basic structural and functional properties. Phylogenetic trees derived from an alignment of kinase domain amino acid sequences (essentially an expanded version of Fig. 1) served as the basis for this classification. Thus, the sole consideration was similarity in kinase domain amino acid sequence. When considered alone, however, this property has been a good indicator of other characteristics held in common by the different members of the family.

Protein kinases whose entire kinase domain amino acid sequence had been published by July 1993 were included in phylogenetic analysis (as well as a few others made available at that time through sequence databases). If a given kinase domain sequence had been determined from more than one species among the vertebrates (i.e., orthologous gene products), only one representative (usually human) was included in the analysis. This policy was not used for the other phyla, however, because of greater divergences between the species and, hence, the sequences. The kinase domain phylogenies were inferred using the principle of maximum parsimony according to the PAUP software package developed by Swofford (41). Minimum-length trees were found using PAUP's 'heuristic' search method with branch swapping by the 'tree bisection-reconnection' strategy. Equal weights were given for all amino acid substitutions. Because multiple minimum-length trees were found, a consensus tree was calculated according to the method of Adams (cited in ref 41) in order to show branching ambiguities.

To accommodate the large numbers of sequences, it was necessary to construct five separate trees. Initially, a skeleton tree of 99 kinases was obtained (**Fig. 3*A***). The skeleton tree included only representative members from each of four large groups of protein kinases, each consisting of multiple related families known from previous work to cluster together in the tree. These four groups are designated: *1)* the AGC group, which includes the cyclic-nucleotide-dependent family (PK<u>A</u> and PK<u>G</u>), the protein kinase C (PK<u>C</u>) family, the β-adrenergic receptor kinase (βARK) family, the ribosomal S6 kinase family, and other close relatives; 2) the CaMK group, which includes the family of protein kinases regulated by calcium/calmodulin, the Snfl/AMPK family, and other close relatives; *3)* the CMGC group, which includes the family of cyclin-dependent kinases, the Erk (MAP) kinase family, the glycogen synthase 3 (GSK3) family, the casein kinase II family, the Clk (Cdk-like kinase) family, and other close relatives; and *4)* the 'conventional' protein-tyrosine kinase (PTK) group. Separate trees (Fig. 3*B-E*) were later obtained for each of the four large kinase groups, and contain all members of the groups whose sequences were available at the time of analysis.


A
AGC Group
PKA-Cα
cPKCα
βARK1
S6K
PVPK1
CaMK Group
CaMK2α
skMLCK
PhKγM
Mro4
PSK-H1
Kin1
Snf1
Polo
Plk
Cdc5
CMGC Group
Cdc2
GSK3α
CK2α
Erk1
Clk
Dsk1
Yak1
Ire1
Cdc7
Cot
TpkA
Ste7
Ste20
Ste11
Byr2
Byr1
Ran1
Nek
Pim1
Wee1
Mik1
NRK
PKR (Tik)
Gcn2
Raf-1
B-Raf
DmRaf
Ctrl
Ctr1
PlyCar
Pelle
Mos
PTK Group
Src
Pctk1
Tk1516
CK1α
CK1β
CK1δ
Hrr25
Yck1
Yck2
C
CaMK2α
CaMK2β
CaMK2γ
DmCaMK2
CaMK2δ
CaMK1
CaMK4
DdMLCK
PsMK1
ScCaMK2-1
ScCaMK2-2
AnCaMK2
skMLCK
smMLCK
Twn
Titin
Mro4
RSK1(Ct)
RSK2(Ct)
PhKγM
PhKγT
CDPK
AtAK1
Kin1
Kin2
SpKin1
p78
Snf1
PSnf1-SKIN11
PSnf1-BKIN12
PSnf1-AKIN10
Nim1
Ycl24
PfCPK
PfPK2
B
PKA-Cα
PKA-Cβ
CePKA
DmPKA-C0
AplC
PKA-Cγ
DmPKA-C1
Sak
DmPKA-C2
DdPKA
ScPKA-Tpk1
ScPKA-Tpk3
ScPKA-Tpk2
DdPK1
PKG-I
DmPKG-G2
DmPKG-G1
cPKCα
cPKCβ
cPKCγ
Apl-I
DmPKC-53Ebr
DmPKC-53Eey
aPKCζ
nPKCε
nPKCη
Apl-II
DmPKC-98F
nPKCδ
nPKCθ
CePKC
ScPKC
Pck1
Pck2
RACα
RACβ
DmRAC
Sgk
βARK1
βARK2
DmGPRK1
RhK
DmGPRK2
Ddk2
Ypk1
Ykr2
Sch9
S6K
RSK1 (Nt)
RSK2 (Nt)
DMPK
NcCot-1
Dbf2
Dbf20
PvPK1
OsG11A
ZmPVK
AtPK5
AtPK64
AtPK7
PsPK5
D
Cdc2
DmCdc2
Cdk2
Cdk3
Cdk5
Cdk6
Cdk4
CfCdc2R
PCTAIRE1
PCTAIRE2
PCTAIRE3
Pho85
DdPRK
PfC2R
PCdc2
SpCdc2
Cdc28
DdCdc2
DmCdc2c
EhC2R
OsC2R
Mo15
Kin28
PITSLRE
Ched
Ctk1
Sgv1
Erk1
Erk2
DmErkA
CaErk1
Kss1
Fus3
Spk1
PErk
Mpk1
Erk3
GSK3α
GSK3β
Sgg
Mck1
CK2α
CeCK2
DmCK2
CK2α'
ScCK2α
ScCK2α'
DdCK2
ZmCK2
TpCK2
KKIALRE
Mak
Sme1
Clk
PskG1
Kns1
PskH2
Yak1
Dsk1

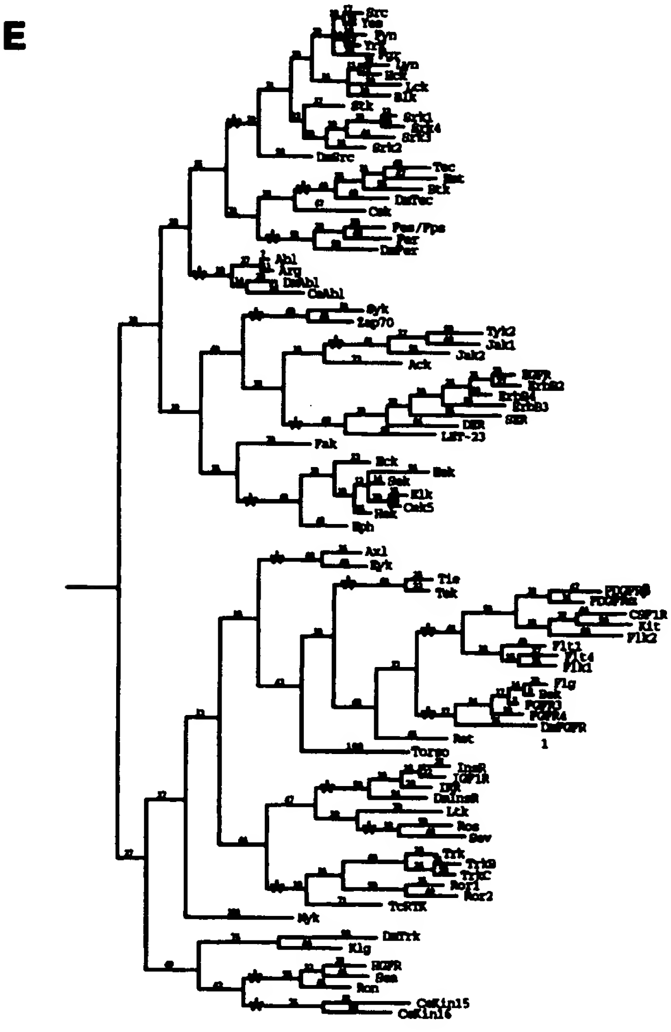


Figure 3. Phylogenetic trees of the eukaryotic protein kinase superfamily inferred from kinase domain amino acid sequence alignments. The abbreviated nomenclature is the same used in Table 1. *A)* 'Skeleton' tree showing 99 protein kinases. Positions of 4 clusters (AGC, CaMK, CMGC, and PTK) containing protein kinases representative of larger groups are indicated in the skeleton tree. *B)* AGC group tree of 59 protein kinases including PKA, PKG, and PKC and other close relatives. *C)* CaMK group tree of 35 protein kinases including the calcium/calmodulin-regulated enzymes. *D)* CMGC group tree of 59 protein kinases including the cyclin-dependent kinases. *E)* PTK group tree of 90 conventional protein-tyrosine kinases. Tree A is unrooted and drawn with Pkn1 and Pkn2 as outgroups. Outgroups of two or more distantly related protein kinases (not shown) were included in the analysis of trees B–E to provide a rooting point. Asterisks (*) in all trees indicate branches leading to defined protein kinase families listed in Table 1. Branch lengths indicate number of amino acid substitutions required to reach hypothetical common ancestors at internal nodes.

It can be reasonably surmised that the protein kinases having closely related catalytic domains, and thus defining a family, represent products of genes that have undergone relatively recent evolutionary separations. Given this, it should come as no surprise that members of a given family tend also to share related functions. This is manifest by similarities in overall structural topology, mode of regulation, and substrate specificity. The details of the common properties exhibited by the members of the various kinase families can best be gleaned from studying the information outlined in the individual entries section of the *Protein Kinase Factsbook* (42). Some of the most salient relationships are discussed below.

The AGC group protein kinases tend to be basic amino acid-directed enzymes, phosphorylating substrates at Ser/Thr residues lying very near Arg and Lys. For the cyclic nucleotide-dependent and ribosomal S6 kinase families, the preferred substrates have basic residues lying in specific positions NH_2-terminal to the phosphate acceptor. Preferred substrates for the PKC and RAC families have basic residues on both the NH_2- and COOH-terminal sides of the acceptor (43). The G-protein-coupled receptor kinases (βARK and RhK) appear to break this rule, however, as they are reported to prefer synthetic peptide substrate residues located within an acidic environment. Little substrate information is available for the other families in this group.

The CaMK group protein kinases also tend to be basic amino acid- directed, and in this regard it is notable that the AGC and CaMK groups fall near one another in the phylogenetic tree. CaMK1, CaMK2, CaMK4, MLCK, CDPK, and AMPK are all reported to prefer substrates with basic residues at specific positions NH_2-terminal to the acceptor site, whereas EF2K and PhK prefer sites with basic residues at both NH_2- and COOH-terminal locations. Many, but not all, of the CaMK group protein kinases are known to be activated by Ca^{2+}/calmodulin binding to a small domain located just COOH-terminal to the catalytic domain, e.g., CaMK1, CaMK2, CaMK4, PhKγ, MLCK, and twitchin. These enzymes and their close relatives are grouped together in a large family within the CaMK group. Also included in this family are a subfamily of plant enzymes (represented by CDPK) that contain an intrinsic calmodulin-like domain that confers Ca^{2+}-dependent activation. The other family within the CaMK group is the Snf1/AMPK family. Within this family, substrate specificity determinant information has been obtained only for the AMP-activated protein kinase, which also shows a requirement for an NH_2-terminal basic residue. The other major category of protein-serine kinases is the CMGC group. For the most part, these are proline-directed enzymes, phosphorylating substrates at sites lying in Pro-rich environments. Available data for Cdc2 and Cdk2 indicate that members of the cyclin-dependent kinase family require phosphate acceptors lying immediately NH_2-terminal to a Pro. A similar requirement is indicated for the Erk (MAP) kinase family. The situation for the GSK3 family is more complicated, but most known acceptor sites lie within Pro-rich regions. The structures of Cdk2 and Erk2 indicate that the pocket for the +1 residue is shallower than in PKA-Cα due to the replacement of Leu205 by an Arg, which is bulkier and precludes binding of the larger hydrophobic amino acids. In addition, the unique secondary amide group of Pro may make special interactions (44). The casein-kinase II family enzymes fail to conform to the proline-directed specificity exhibited by the other major families of this group, showing instead a strong preference for Ser residues located NH_2-terminal to a cluster of acidic residues. The CMGC group protein kinases have larger-than-average kinase domains due to insertions between subdomains X and XI, whose functional significance is unknown.

The conventional protein-trosine kinase group includes a large number of enzymes with quite closely related kinase domains that specifically phosphorylate on Tyr residues (i.e., they cannot phosphorylate Ser or Thr). These enzymes, first recognized among retroviral oncoproteins, have been found only in metazoan cells where they are widely recognized for their roles in transducing growth and differentiation signals. Included in this group are more than a dozen distinct receptor families made up of membrane-spanning molecules that share similar overall structural topologies, and nine nonreceptor families also composed of structurally similar molecules. The specificity determinants surrounding the Tyr phosphoacceptor sites have yet to be firmly established for these enzymes, but Glu residues either on the NH_2- or COOH-terminal side of the acceptor are often preferred. This group is labeled "conventional" to distinguish it from other protein kinases (including Spk1, Clk, the MEK/Ste7 family members, Weel/Mik1, ActRII, Hrr25, Esk, and Spl A/DPyk2) reported to exhibit a dual specificity, that is, being capable of phosphorylating both Tyr and Ser/Thr residues (45). However, in most cases dual specificity has been observed only for autophosphorylation reactions in vitro, and the only dual specificity protein kinases that are known to be able to phosphorylate a substrate on Ser/Thr and Tyr are members of the MEK family. Considered as a group, these dual-specificity protein kinases are not particularly closely related to the conventional PTKs. Indeed, they seem to map throughout the phylogenetic tree (45), suggesting that the ability to autophosphorylate on Tyr may have had many independent origins during the evolutionary history of the superfamily.

The protein kinases falling outside the four major groups are a mixed bag. Although the individual members within the defined families found in this "other" category clearly are related to one another through both structure and function, it is difficult to make broader generalizations that could group anv of these families together into a larger category. As far as substrate specificity determinants go, little is known about most "other" category protein kinases, due primarily to their rather recent discovery and the paucity of known physiological substrates. The casein kinase I family members, however, have been shown to prefer Ser/Thr residues located COOH-terminal to a phosphoserine or phosphothreonine, although a stretch of acidic residues may substitute. Also, the family of protein kinases involved in translational control (HRI, PKR/Tik, Gcn2) appear to be basic amino acid-directed enzymes preferring Ser residues lying NH_2- terminal to an Arg. Finally, as mentioned previously, the MEK/Ste7 family protein kinases and Weel/Mik1 protein kinases exhibit a dual specificity.

Although this classification is based solely on catalytic domain sequences, members of families defined by this means are usually closely related in regions lying outside the cataytic domains and in many cases have been shown to possess very similar functions. Thus, intercalation of newly discovered protein kinases into this classification should allow one to make useful predictions about the functions of such enzymes.

FUTURE PROSPECTS

The rate of protein kinase discovery still shows no signs of abating. In addition to the continuing successes of homology-based approaches, genomic sequencing projects are beginning to make significant contributions. For instance, the sequences of two entire budding yeast chromosomes (46, 47) and a ~2 Mb stretch of *C. elegans* chromosome III (48) have revealed a number of new putative protein kinase genes. As genome sequencing projects gather speed, the number of new protein kinase genes discovered in this way will undoubtedly mushroom. This explosion of sequence data is making it increasingly difficult to manage protein kinase databases of the sort described here. Programs designed to align and derive relatedness trees are currently unable to handle the large number of available kinase domain sequences. New data handling programs will have to be developed to cope with large numbers of sequences like those of the eukaryotic protein kinase superfamily.

Protein kinase catalytic domain structures will continue to be solved. The first structure of a conventional protein-tyrosine kinase will be available shortly (see "Note added in proof"), and this should reveal how Tyr is selected as an acceptor amino acid vs. Ser/Thr. Such structures will enable comparative analysis to be carried out at the 3-dimensional level, and allow predictions of structures from primary sequences. Structural comparisons of catalytic domains with bound peptide substrates will also provide insights into substrate specificity. Most protein kinases show some degree of primary sequence specificity, and new methods are being developed to determine consensus sequence specificities for individual protein kinases (44). With such consensus information the structural basis for the binding of a preferred peptide sequence to the cognate substrate binding site can then be deduced. In the future, it may be possible to model the 3-dimensional structure of a novel protein kinase catalytic domain with sufficient accuracy to be able to deduce the preferred primary sequence surrounding the hydroxyamino acid it phosphorylates, which in turn will allow one to predict what proteins might be its substrates from the increasingly complete database of protein sequences. FJ

Note added in proof: The crystal structure of the tyrosine kinase domain of the insulin receptor has now appeared (Hubbard, S. R., Wei, L., Ellis, L., and Hendrickson, W. A. (1994) *Nature* **372**, 746–754).

REFERENCES

1. Hanks, S. K., Quinn, A. M., and Hunter, T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42-52
2. Hanks, S. K. (1991) Eukaryotic protein kinases. *Curr. Opin. Struct. Biol.* **1**, 369-3833.
3. Hanks, S. K., and Quinn, A. M. (1991) Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* **200**, 38-62
4. Hunter, T. (1987) A thousand and one protein kinases. *Cell* **50**, 823-8295.
5. Hunter, T. (1994) 1001 protein kinases redux: towards 2000. *Seminars Cell Biol.* In press
6. Muñoz-Dorado, J., Inouye, S., and Inouye, M. (1991) A gene encoding a protein serine/threonine kinase is required for normal development of *M. xanthus*, a gram-negative bacterium. *Cell* **67**, 995-1006
7. Galyov, E. E., Hakansson, S., Forsberg, A., and Wolf-Watz, H. (1993) A secreted protein kinase of *Yersinia pseudotuberculosis* is an indispensable virulence determinant. *Nature* **361**, 730-732
8. Hoekstra, M. F., DeMaggio, A. J., and Dhillon, N. (1991) Genetically identified protein kinases in yeast. Part 1: transcription, translation, transport and mating. *Trends Genet.* **7**, 256-261
9. Alex, L. A., Simon, M. J. (1994) Protein histidine kinases and signal transduction in prokaryotes and eukaryotes. *Trends Genet.* **10**, 133-136
10. Chang, C., Kwok, S. F., Bleecker, A. B., and Meyerowitz, E. M. (1993) *Arabidopsis* ethylene-response gene *ETR1*: similarity of product to two-component regulators. *Science* **262**, 539-544
11. Ota, I. M., and Varshavsky, A. (1993) A yeast protein similar to bacterial two-component regulators. *Science* **262**, 566-569
12. Maeda, T., Wurgler-Murphy, S. M., and Saito, H. (1994) A two-component system that regulates an osmosensing MAP kinase cascade in yeast. *Nature* **369**, 242-245
13. Popov, K. M., Zhao, Y., Shimomura, Y., Kuntz, M. J., and Harris, R. A. (1992) Branched-chain α-ketoacid dehydrogenase kinase. *J. Biol. Chem.* **267**, 13127-13130
14. Popov, K. M., Kedishvili, N. Y., Zhao, Y., Shimomura, Y., Crabb, D. W., and Harris, R. A. (1993) Primary structure of pyruvate dehydrogenase kinase establishes a new family of eukaryotic protein kinases. *J. Biol. Chem.* **268**, 26602-26606
15. Maru, Y., and Witte, O. N. (1991) The BCR gene encodes a novel serine/threonine kinase activity within a single exon. *Cell* **67**, 459-468
16. Beeler, J. F., LaRochelle, W. J., Chedid, M., Tronick, S. R., and Aaronson, S. A. (1994) Prokaryotic expression cloning of a novel human tyrosine kinase. *Mol. Cell. Biol.* **14**, 982-988
17. Huang, J. M., Wei, Y. F., Kim, Y. H., Osterberg, L., and Matthews, H. R. (1991) Purification of a protein histidine kinase from the yeast *Saccharomyces cerevisiae*. The first member of this class of protein kinases. *J. Biol. Chem.* **266** 9023-9031
18. Stock, J. B., Ninfa, A. J., and Stock, A. M. (1989) Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiol. Rev.* **53**, 450-490
19. Cozzone, A. J. (1993) ATP-dependent protein kinases in bacteria. *J. Cell. Biochem.* **51**, 7-13
20. Saier, M. H. (1993) Introduction: protein phosphorylation and signal transduction in bacteria. *J. Cell. Biochem.* **51**, 1-6
21. Reizer, J., Romano, A. H., and Deutscher, J. (1993) The role of phosphorylation of HPr, a phosphocarrier protein of the phosphotransfer system, in the regulation of carbon metabolism in gram-positive bacteria. *J. Cell. Biochem.* *4751*, 19-24
22. LaPorte, D. C. (1993) Isocitrate dehydrogenase phosphorylation cycle: regulation and enzymology. *J. Cell. Biochem.* **51**, 14-18
23. Muñoz-Dorado, J., Inouye, S., and Inouye, M. (1993) Eukaryotic-like protein serine/threonine kinases in *Myxococcus xanthus*, a developmental bacterium exhibiting social behavior. *J. Cell. Biochem.* **51**, 29-33
24. Zhang, C.-C. (1993) A gene encoding a protein-related to eukaryotic protein kinases from the filamentous heterocystous cyanobacterium *Anabena* PCC7120. *Proc. Natl. Acad. Sci. USA* **90**, 11840-11844
25. Knighton, D. R., Zheng, J., Ten Eyck, L. F., Xuong, N.-H., Taylor, S. S., and Sowadski, J. M. (1991) Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 414-420

26 Knighton, D. R., Zheng, J., Ten Eyck, L. F., Ashford, V. A., Xuong, N.-H., Taylor, S. S., and Sowadski, J. M. (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 407-420

27. Bossemeyer, D., Engh, R. A., Kinzel, V., Ponstingl, H., and Huber, R. (1993) Phosphotransferase and substrate binding mechanism of the cAMP-dependent protein kinase catalytic subunit from porcine heart as deduced from the 2.0 Å structure of the complex with Mn^{2+} adenyl imidodiphosphate and inhibitor peptide PKI(5-24). *EMBO J.* **12**, 849-859
28. Zheng, J., Knighton, D. R., ten Eyck, L. F., Karlsson, R., Xuong, N., Taylor, S. S., and Sowadski, J. M. (1993) Crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MgATP and peptide inhibitor. *Biochemistry* **32**, 2154-2161
29. De Bondt, H. L., Rosenblatt, J., Jancarik, J., Jones, H. D., Morgan, D. O., and Kim, S. (1993) Crystal structure of cyclin-dependent kinase 2. *Nature* **363**, 595-602
30. Zhang, F., Strand, A., Robbins, D., Cobb, M. H., and Goldsmith, E. J. (1994) Atomic structure of the MAP kinase ERK2 at 2.3 Å resolution. *Nature* **367**, 704-711
31. Hu, S.-H., Parker, M. W., Lei, J. Y., Wilce, M. C. J., Benian, G. M., and Kemp, B. E. (1994) Insights into autoregulation from the crystal structure of twitchin kinase. *Nature* **369**, 581-584
32. Carmel, G., Leichus, B., Cheng, X., Patterson, S. D., Mirza, U., Chait, B. T., and Kuret, J. (1994) Expression, purification, crystallization, and preliminary X-ray analysis of casein kinase-1 from *Schizosaccharomyces pombe*. *J. Biol. Chem.* **269**, 7304-7309
33. Zheng, J., Knighton, D. R., Xuong, N. H., Taylor, S. S., Sowadski, J. M., and Ten Eyck, L. F. (1993) Crystal structures of the myristylated catalytic subunit of cAMP-dependent protein kinase reveal open and closed conformations. *Protein Sci.* **2**, 1559nd573
34. Knighton, D. R., Cadena, D. L., Zheng, J., Ten Eyck, L. F., Taylor, S. S., Sowadski, J. M., and Gill, G. N. (1993) Structural features that specify tyrosine kinase activity deduced from homology modeling of the epidermal growth factor receptor. *Proc. Natl. Acad. Sci. USA* **90**, 5001-5005
35. Taylor, S. S., Knighton, D. R., Zheng, J., Ten Eyck, L. F., and Sowadski, J. M. (1992) Structural framework of the protein kinase family. *Annu. Rev. Cell Biol.* **8**, 429-462
36. Taylor, S. S., Zheng, J., Radzio-Andzelm, E., Knighton, D. R., Ten Eyck, L. F., Sowadski, J. M., Herberg, F. W., and Yonemoto, W. (1993) cAMP-dependent protein kinase defines a family of enzymes. *Phil. Trans. R. Soc. London B* **340**, 315-324
37. Madhusudan, A., Trafny, E. A., Xuong, N. H., Adams, J. A., Ten Eyck, L. F., Taylor, S. S., and Sowadski, J. M. (1994) cAMP-dependent protein kinase: crystallographic insights into substrate recognition and phosphotransfer. *Protein Sci.* **3**, 176-187
38. Taylor, S. S., Radzio-Andzelm, E. (1994) Three protein kinase structures define a common motif. *Structure* **2**, 345-355
39. Steinberg, R. A., Cauthron, R. D., Symcox, M. M., and Shuntoh, H. (1993) Autoactivation of catalytic (Cα) subunit of cyclic AMP-dependent protein kinase by phosphorylation at threonine 197. *Mol. Cell. Biol.* **13**, 2332-2341
40. Veron, M., Radzio-Andzelm, E., Tsigelny, I., Ten Eyck, L. F., and Taylor, S. S. (1993) A conserved helix motif complements the protein kinase core. *Proc. Natl. Acad. Sci. USA* **90**, 10618-10622
41. Swofford, D. (1991) PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1. Illinois Natural History Survey, Champaign, Illinois
42. Hardie, D. G., and Hanks, S. K. (1995) The Protein Kinase Factsbook, Academic Press, London
43. Pearson, R. B., Kemp, B. E. (1991) Protein kinase phosphorylation site sequences and consensus specificity motifs: tabulations. *Meth. Enzymol.* **200**, 62-81
44. Songyang, Z., Blechner, S., Piwnica-Worms, H., and Cantley, L. C. (1994) A novel oriented peptide library technique for determining optimal substrates of protein kinases. *Curr. Biol.* In press
45. Lindberg, R. A., Quinn, A. M., and Hunter, T. (1992) Dual-specificity protein kinases: will any hydroxyl do? *Trends Biochem. Sci.* **17**, 114-119
46. Koonin, E. V., Bork, P., and Sander, C. (1994) Yeast chromosome III: new gene functions *EMBO J.* **13**, 493-503
47. Johnston, M., Andrews, S., Brinkman, R., Cooper, J., Ding, H., Dover, J., Du, Z., Pavello, A., Fulton, L., Gattung, S., et al. (1994) Complete nucleotide sequence of Saccharomyces cerevisiae chromosome VIII. *Science* **265**, 2077-2082
48. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J ., Burton, J., Connell, M., Copsey, T., Cooper, J., et al. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*.. *Nature* **368**, 32-38

Protein Kinases and Phosphatases in the *Drosophila* Genome

Deborah K. Morrison, Monica S. Murakami, and Vaughn Cleghon

Regulation of Cell Growth Laboratory, National Cancer Institute, Frederick, Maryland 21702

The reversible phosphorylation of proteins on serine, threonine, and tyrosine residues represents a fundamental strategy used by eukaryotic organisms to regulate a host of biological functions, including DNA replication, cell cycle progression, energy metabolism, and cell growth and differentiation. Levels of cellular protein phosphorylation are modulated both by protein kinases and phosphatases. Although the importance of kinases in this process has long been recognized, an appreciation for the complex and fundamental role of phosphatases is more recent. Through extensive biochemical and genetic analysis, we now know that pathways are not simply switched on with kinases and off with phosphatases. Rather, it is the balance of phosphorylation that is often critical. Protein phosphorylation can regulate enzyme function, mediate protein–protein interactions, alter subcellular localization, and control protein stability. Furthermore, kinases and phosphatases may work together to modulate the strength of a signal. Adding further complexity to this picture is the fact that both kinases and phosphatases can function in signaling networks where multiple kinases and phosphatases contribute to the outcome of a pathway. To fully understand this complex and essential regulatory process, the kinases and phosphatases mediating the changes in cellular phosphorylation must be identified and characterized.

A variety of approaches, including biochemical purification, gene isolation by homology, and genetic screens, have been successfully used for the identification of putative protein kinases and phosphatases. Now, the genomic sequencing of organisms promises to be a major contributor to this field. Valuable insight into these important enzymes has already emerged from the analysis of the yeast and worm genomes. In particular, genomic sequencing of *Saccharomyces cerevisiae* and *Caenorhabditis elegans* has revealed the kinase and phosphatase gene families that have arisen during the evolution of multicellular eukaryotes (Plowman et al., 1999). With the recent determination of the *Drosophila* sequence, we can now survey the genome of a second multicellular eukaryote for its repertoire of kinases and phosphatases. In this review, we will present our findings on the protein kinase and phosphatase gene families identified in the fly, together with an examination of the kinase/phosphatase signaling pathways functioning in flies, worms, and humans.

Identification and Classification of Drosophila Protein Kinases and Phosphatases

Our survey of *Drosophila* protein kinases and phosphatases is based on the total set of predicted proteins that were identified in the *Drosophila* genome using automated gene predictor methods (Adams et al., 2000; available at http://www.celera.com). The 13,601 predicted fly proteins were surveyed for overall homology with known kinase and phosphatase sequences using BLASTP, and for the presence of polypeptide motifs using BLOCKS and InterPro databases (Rubin et al., 2000). Putative kinases and phosphatases identified by these means were further classified based on the presence of diagnostic amino acid residues in conserved motifs and by sequence similarities extending beyond conserved catalytic domains. Table I summarizes our survey of the *Drosophila* protein kinases and phosphatases. It is important to realize that this analysis represents the first tabulation of these enzymes in *Drosophila* and will be subject to revision as gaps in the genomic sequence are closed and methods for predicting and analyzing genes are improved. In particular, it is known that the Genie and Genscan programs used to annotate the fly genomic sequence make systematic errors with respect to intron–exon boundaries and gene borders, leading us to conclude that some kinase and phosphatase proteins may have been missed by these programs (Reese et al., 2000). These caveats notwithstanding, 251 kinases and 86 phosphatases were identified by our analysis of the predicted *Drosophila* protein set. Remarkably, more than half of these molecules had gone undetected in eight decades of *Drosophila* research.

Protein Kinases

Eukaryotic protein kinases are enzymes that catalyze the transfer of phosphate from ATP or GTP onto serine, threonine, or tyrosine residues of their appropriate substrates. They comprise a single protein superfamily having a common catalytic structure. However, these enzymes can be subdivided into distinct groups based on their structural and functional properties (Hanks and Hunter, 1995).

AGC Family

The AGC serine/threonine kinases function in many intracellular signaling pathways and were first classified based on their tendency to phosphorylate sites surrounded by basic amino acids. *Drosophila* contains ~30 AGC kinases, including members of the cyclic nucleotide-dependent ki-

Address correspondence to Deborah Morrison, NCI-FCRDC, P.O. Box B, Frederick, MD 21702. Tel.: (301) 846-1733. Fax: (301) 846-1666. E-mail: morrisod@nciaxp.ncifcrf.gov

Vaughn Cleghon's current address is Beatson Institute CRC, Glasgow G61 1BD, UK. E-mail: v.cleghon@beatson.gla.ac.uk

Supplemental Material can be found at:
http://www.jcb.org/cgi/content/full/150/2/F57/DC1

Table I. Summary of Protein Kinases and Phosphatases in Flies, Worms, and Humans

Group	Fly	Worm*	Humans*
Protein kinase			
AGC	30 (8)	30	100
CaMK	25 (13)	32	83
CKI	8 (6)	87	5
CMGC	24 (7)	42	62
STE	21 (12)	28	63
PTK	32 (8)	92	100
OPK	56 (28)	62	163
Atypical	3 (2)	4	11
Fragment/unknown	18		
Protein kinase like			
Gcyc	11 (6)	26	8
PIK	13 (8)	12	20
DAG	8 (5)	7	8
Choline K	2 (1)	7	2
Phosphatase			
STP	28 (14)	65	21
RPTP, CPTP, LMW-PTP	20 (12)	83	47
DSP	18 (11)	26	51
IPP	20 (18)	11	7

Fly numbers in parentheses represent the proteins newly identified by the fly genome project.
*These numbers are taken from the review by Plowman et al. (1999).

nases, protein kinase C (PKC),[1] AKT, NDR, MNK, MAST, ribosomal S6 kinase, and G protein–coupled receptor kinase families. The majority of the fly AGC kinases had been identified previously by molecular and genetic analysis; however, eight members were uncovered in the fly genome project. Interestingly, four of the new genes encode PKC or PKC-related proteins, including the first atypical PKC isoforms identified in *Drosophila.* Also identified by the fly genome project were additional PKA and PKG proteins, as well as kinases related to mammalian MAST205 and Citron.

CaMK Family

The CaMK serine/threonine kinases also tend to have substrate recognition motifs containing basic amino acids, and some but not all members of this family are regulated by calcium or calmodulin. Approximately 25 CaMKs are present in *Drosophila*, including representatives of the calcium/calmodulin-regulated kinase, SNF1/AMP-dependent kinase, EMK, CHK2, myosin light chain kinase (MLCK), phosphorylase kinase, death-associated protein kinase, and MAPKAP kinase families (the last four of which are found in *C. elegans* but not yeast). Like worms, flies do not encode a complete ortholog of the mammalian Trio kinase, but do have a protein that is related to the entire Trio regulatory domain. CaMK members revealed by the fly genome project include proteins related to calcium/calmodulin-regulated kinases, MLCK, EMK, and mammalian DRAK1. Of the 13 newly identified CaMKs, 6 belong to the EMK family, making this the largest CaMK group in flies. Mammalian and *C. elegans* EMK proteins have been implicated in the regulation of cell polarity and microtubule stability (Drewes et al., 1998).

Casein Kinase I Family

The casein kinase I (CKI) proteins originally were characterized as ubiquitous serine/threonine kinases with a preference for acidic substrates such as casein. Although members of this family were among the first kinases purified, elucidating their function and regulation has been difficult. Recently, however, CKI isoforms have been found to play a role in DNA repair and cell division (Gross and Anderson, 1998), in the Wnt signaling pathway (Peters et al., 1999), and in circadian rhythm regulation (Lowrey et al., 2000). *Drosophila* contains at least eight CKI proteins, only two of which were known previously. Intriguingly, CKI is one of the kinase families that is significantly expanded in the worm, with 87 members identified in *C. elegans* (Plowman et al., 1999). The biological significance of the worm-specific expansion is currently unknown.

CMGC Family

CMGC family members are primarily proline-directed serine/threonine kinases. The major subfamilies of this group play key roles in cell cycle regulation and intracellular signal transduction, and, not surprisingly, are conserved from yeast to humans. Approximately 24 CMGC kinases are found in *Drosophila*, including members of the cyclin-dependent kinase (CDK), CDC-like kinase (CLK), glycogen synthase kinase 3 (GSK3), and MAPK families. Although extensive genetic analysis had revealed many of the *Drosophila* CMGC kinases, seven novel proteins were uncovered by the fly genome project. These include additional CDK (CDK7-like, CDC2-related KKIALRE, CHED-related), GSK3, and MAPK (ERK7) members, as well as an RCK family member (MAK). Also uncovered in the fly genome were proteins related to the MP1 and JIP-1 scaffolding proteins. These molecules function to localize MAPK proteins with their upstream activators and provide signaling specificity (Whitmarsh and Davis, 1998). Although MAPK scaffolding proteins are present in yeast, they are structurally different from the ones found in flies, worms, and mammals, perhaps indicating the evolution of these molecules in multicellular eukaryotes.

STE Family

The STE family is composed of the STE7 (MEK), STE11 (MEKK), and STE20 (MEKKK) kinases that function upstream of MAPK proteins. *Drosophila* contains ~21 members of this family, only 9 of which were known previously. Remarkably, 9 members of the PAK/STE20 group were uncovered by the fly genome project, including proteins related to mammalian PAK3, GLK1, NIK, MST2, STLK3, TAO1, and CDC7. Although PAK proteins containing PH domains are found in yeast (Sells et al., 1999), no PH-domain-containing PAKs have been identified in higher eukaryotes and none are present in *Drosophila.* MEKK- and NEK-related kinases were also revealed by the genome project. It is worth noting that even with the discovery of additional MEK and MAPK proteins in the fly, *C. elegans*

[1] *Abbreviations used in this paper:* CDK, cyclin-dependent kinase; CKI, casein kinase I; CTK, cytoplasmic tyrosine kinase; DSP, dual specificity phosphatase; LMW, low molecular weight; MKP, MAPK phosphatase; PKC, protein kinase C; PTP, protein tyrosine phosphatase; RTK, receptor tyrosine kinase; STP, serine/threonine protein phosphatase.

contains over twice as many of these kinases, suggesting an expansion of MAPK signaling modules in the worm.

PTK Family

The PTK group consists of receptor (RTK) and cytoplasmic (CTK) tyrosine kinases. Although yeasts contain no conventional PTKs, 92 have been identified in the worm and ~32 are present in the fly. A major function of PTKs is in intercellular communication, perhaps explaining why these enzymes have only been identified in multicellular eukaryotes. In comparison to *Drosophila*, the much larger number of PTKs found in *C. elegans* is due primarily to expansions of the worm-specific Kin-15/16 RTK and FER CTK families. The majority of the fly PTKs had been identified previously by genetic approaches, reflecting the involvement of these proteins in critical growth and developmental pathways. RTKs encoded in the fly genome include the fly-specific Torso and Sevenless kinases, as well as kinases related in sequence if not function to the mammalian EGFR, FGFR, insulin receptor, EPH, RET, ROR, RYK, ALK, and TRK kinases. Of the five newly identified RTKs, two are related to mammalian PDGFR/VEGFR, two are DDR receptors, and one shares homology with FGFR1. In the CTK group, fly members include the JAK, FAK, SYK/SHARK, ACK, ABL, and FPS kinases. Of the newly identified CTKs, one is related to mammalian ACK2 and one is an ortholog of CSK, a kinase that negatively regulates the activity of mammalian SRC kinases. Interestingly, several members of the PTK class are not found in worms, including representatives of the SYK, JAK, TRK, and RET families.

OPK Group

This group is comprised of other protein kinase (OPK) families that do not belong to the six major groups described above. It is the largest class of kinases found in flies and consists of both serine/threonine and dual specificity kinases. Approximately 56 of these enzymes are present in the fly genome, only half of which were known previously. Representatives of this group are extremely diverse and include members of the following families: Aurora, BUB1, CHK1, DYRK, WEE-1, PLK, EIF2, TGFβ, and activin receptor, TAK, IKK kinases, CKII, and RAF kinase. Notable in the novel group are additional BUB1 and TAK members and enzymes related to *C. elegans* UNC 51 and mammalian ALK3, DLK, GAK, MLK2, SRPK, IRE, ILK, TLK1, LIM-domain kinase, and LKB1/Peutz-Jeghers kinase.

Atypical, Lipid, and Unknown Kinases

Several protein groups that are structurally related to the eukaryotic protein kinases are also found in the *Drosophila* genome. These include the atypical kinases, guanylyl cyclases, and the eukaryotic lipid kinases. Flies contains at least three atypical kinase members, pyruvate dehydrogenase kinase, A6, and a newly identified BCR protein. Although worms lack BCR, they do contain a protein related to the atypical *Dictyostelium* myosin heavy chain kinase, which appears to be missing in flies. Also absent in both *Drosophila* and *C. elegans* are representatives of the classical prokaryotic histidine kinases. In the lipid kinase group, *Drosophila* encodes at least 8 diacylglycerol kinases, 2 choline/ethanolamine kinases, and 13 phophatidylinositol kinases (PI3-, PI4-, PIP5,- and PIP3-related kinases), the majority of which were unknown previously. In mammalian cells, members of the PIP3-related kinase family participate in the cellular response to DNA damage and have authentic protein kinase activity (for review see Fruman et al., 1998). The fly genome project has revealed three kinases of this group, namely ATM, FRAP-related protein (FRP), and FRAP/TOR; however, as is true for worms, flies do not contain a DNA-PK. Finally, ~18 proteins were identified that represent either partial kinase fragments or kinases with no significant homology to the groups listed above. Since errors have been identified in the transcript annotation of several protein kinases, such as the DDR receptors, Citron, and a PKC isoform, some of the partial kinase sequences may represent intact enzymes that have been improperly annotated. Further analysis will be required to confirm their identity.

Protein Phosphatases

Unlike protein kinases, which share a common catalytic structure, protein phosphatases have different basic structures, use distinct catalytic mechanisms, and comprise at least three separate protein families. Phosphatases are typically classified into two main groups, the serine/threonine protein phosphatases (STPs) and protein tyrosine phosphatases (PTPs).

STPs

STPs can be subdivided into the PPP and PPM families based on distinct amino acid sequences and crystal structures (for review see Cohen, 1997). Both families are widely distributed across phyla with representatives found in yeast, flies, worms, and mammals. Before the *Drosophila* sequencing project, almost all known fly STPs had been identified by molecular cloning approaches. Very few STPs have been isolated by genetic analysis, indicating that shared substrate specificity and/or functional redundancy may have prevented the recovery of such mutants. *Drosophila* contains ~28 STPs, whereas >65 are encoded in the *C. elegans* genome. The increased number of worm STPs appears to be due to an expansion of the PPP family. Members of the PPP family, such as PP1, PP2A, and PP2B, have been implicated in numerous biological processes and signal transduction pathways. The diverse functions of this family are accomplished by a relatively small number of highly conserved catalytic subunits that complex with a wide variety of regulatory proteins, thus targeting the enzyme to specific intracellular locations and substrates. The *Drosophila* genome encodes ~17 PPP catalytic proteins, 8 PP1-related enzymes (including PP1s, PPN, and PPY), 4 PP2A members (including PP2A, PP4, and PPV), 3 PP2B-like molecules, and 2 PP5 proteins. Additional PPP catalytic subunits uncovered by the fly genome project include members of the PP1, PP4, and PP2B groups. In regard to PPP regulatory subunits, *Drosophila* contains at least 3 PP1, 5 PP2A, and 2 PP2B proteins. However, because the regulatory subunits are so diverse, these numbers are likely to be low.

The PPM family includes PP2C and mitochondrial pyruvate dehydrogenase phosphatase. Due to their highly divergent primary sequences, few PPM members have been isolated by homology-based methods and none have been identified by genetic analysis. The only *Drosophila* PP2C protein that had been previously known was identified by genomic walking (Dick et al., 1997). Remarkably, the genome project has uncovered at least 11 new PP2C-related sequences, including one that closely resembles pyruvate dehydrogenase phosphatase. The biological function of the PPM family has been difficult to assess in mammalian cells due to the lack of specific inhibitors that target these enzymes. Recently, however, a PP2C protein has been found to dephosphorylate CDC2 on Thr161 in yeast (Cheng et al., 1999). Whether any of the PP2Cs perform a similar function in *Drosophila* waits to be determined.

PTPs

PTPs are found in all eukaryotic organisms, and are defined by the catalytic signature motif Cys-X5-Arg (for review see Neel and Tonks, 1997). The PTP superfamily consists of classical PTPs (RPTP, CPTP), dual specificity phosphatases (DSPs), and low molecular weight (LMW) PTPs. Approximately 38 PTPs are encoded in the fly genome, including representatives of each class. Again, many more PTPs are found in the worm (109 total). It is interesting to note that the expansion of serine/threonine and tyrosine kinase families in worms has been accompanied by a corresponding expansion of both serine/threonine and tyrosine phosphatases.

Members of the classical PTP family contain a conserved catalytic domain that is often fused to a large noncatalytic region. The PTP noncatalytic domains are quite diverse and can function to regulate enzyme activity and/or mediate protein interactions. Like PTKs, classical PTPs can be divided into two groups, receptor PTPs (RPTPs) and cytoplasmic PTPs (CPTPs). Genetic studies in *Drosophila* have been instrumental to our understanding of both groups. In particular, experiments in the fly were among the first to demonstrate the involvement of RPTPs in neuronal axon guidance (for review see Desai et al., 1997; den Hertog, 1999). *Drosophila* encodes ~8 RPTKs, at least 5 of which function in this capacity. Of the newly identified RPTPs, one is related to mammalian RPTP-κ and two share homology with RPTP-X/1A2, a type 1 transmembrane PTP implicated in nervous system development and insulin-mediated pancreatic function. In regard to the CPTP class, *Drosophila* studies on the CSW phosphatase were pivotal in demonstrating that a CPTP could function as a positive effector of cell signaling (Perkins et al., 1992). CSW is a member of the SH2-domain containing PTPs (SHP subclass). Mammals are known to have at least two SHPs, whereas no additional SHP proteins were found in *Drosophila*, indicating that flies, like worms, possess a single SHP molecule. Overall the fly genome encodes at least 5 CPTPs, namely CSW, PTP-ER, and newly identified CPTPs related to the mammalian MEG1, MEG2, and PTPD1 phosphatases. Finally, *Drosophila* contains four additional PTP-related proteins which are either difficult to classify or represent incomplete phosphatase fragments.

DSPs are a diverse collection of phosphatase subgroups that share little sequence homology outside of the conserved Cys-X5-Arg motif with other DSP subgroups or with members of the larger PTP family. DSPs were originally characterized by their ability to dephosphorylate both serine/threonine and tyrosine residues; however, some of the DSP subgroups, namely PTEN and myotubularin, also possess lipid phosphatase activity (Maehama and Dixon, 1999). Approximately 18 DSPs are found in *Drosophila*, including representatives of the MAPK phosphatase (MKP), PTEN, nuclear prenylated PRL, myotubularin, PIR1, CDC14, and CDC25 phosphatase groups. Of the nine DSPs uncovered by the fly genome project, six belong to the MKP group, a remarkable finding considering the extraordinary effort spent studying MAPK pathways in *Drosophila*. Only Puckered, a negative regulator of the JNK pathway, previously had been identified by genetic techniques (Martin-Blanco et al., 1998). The failure of the new MKPs to be uncovered by genetic analysis may indicate that they participate in MAPK pathways controlling subtle or unappreciated phenotypes. Alternatively, their functions may have been obscured by redundancy within the MKP group or with other phosphatases. Additional DSPs revealed by the genome project include enzymes related to CDC14 and myotubularin. Interestingly, flies also contain three myotubularin-related sequences that lack the active site Cys and Arg residues. As has been suggested for similar mammalian myotubularin-related molecules, these proteins may function as antiphosphatases by binding to and protecting substrates from dephosphorylation by myotubularin or related phosphatase (Hunter, 1998; for review see Laporte et al., 1998).

LMW-PTPs are ~150-amino acid residue cytoplasmic enzymes that have been shown to possess tyrosine phosphatase activity (Ostanin et al., 1995). Other than a strictly conserved Cys-X5-Arg catalytic motif, LMW-PTPs bear little resemblance to the other PTP members. Mammalian LMW-PTPs have been implicated to function in EPH (Stein et al., 1998) and PDGF receptor signaling (Chiarugi et al., 2000); however, much remains to be learned regarding the biological activity of these enzymes. Although two putative LMW-PTPs are revealed by the *Drosophila* genome project, both predicted proteins are larger than would be expected (424 and 250 amino acids, respectively). The smaller protein contains a complete LMW-PTP domain but lacks the conserved Arg residue in the catalytic motif. Intriguingly, the larger protein has two complete LMW-PTP domains. Although the first domain has a mutation in the active site Cys residue and is likely to be inactive, the second domain contains an intact PTP catalytic motif and presumably has catalytic activity. If this protein is made in vivo, it would represent a new type of LMW-PTP having a tandem catalytic domain structure similar to that observed in many RPTPs. Whether this molecule is an authentic LMW-PTP and whether it has a human counterpart remains to be determined.

Lipid Phosphatases

Lipid inositol phosphatases play an important role in mediating the intracellular balance of second messenger phospholipids. *Drosophila* encodes approximately 20 inositol phosphatases (IPP), only 2 of which were known previously. Six inositol-1,4,5-triphosphase phosphatase–like enzymes are contained in the fly genome; yet as is true for worms, no ortholog of the mammalian SH2-domain–containing inositol 5′ phosphatase (SHIP) appears to be present. *Drosophila* does encode eight PPAP enzymes, which dephosphorylate phosphatidic acid to generate diacylglycerol. The prototype member of this class, Wunen, was first identified in a genetic screen for factors controlling germ cell migration in the early *Drosophila* embryo (Zhang et al., 1996). Related proteins were subsequently identified in yeast, worms, and mammals. Remarkably, the fly genome project reveals seven additional Wunen-like phosphatases. Also uncovered by the genome project are six members of the inositol monophosphate phosphatase (IMP) group. Both the Wunen-like and inositol monophosphate phosphatases are characterized by small tandem gene arrangements, suggesting a limited expansion of these phosphatase families in *Drosophila*. The large number of newly identified inositol phosphatases underscores the hitherto unappreciated importance of lipid phosphoregulation in the fly.

Comparative Analysis of Phosphorylation-dependent Signaling Pathways

With the completion of both the *Drosophila* and *C. elegans* genome projects, together with our current knowledge of mammalian signaling pathways, we can begin to draw conclusions regarding the regulatory complexity of protein phosphorylation mechanisms across the evolutionary spectrum. For example, in flies, worms, and humans, there is a high degree of structural and functional conservation between the components of the RTK and stress-activated signaling pathways, with the major difference being the number of isoforms present for individual pathway members. In higher organisms, the number of isoforms is increased, presumably providing greater potential for tissue- or stage-specific functions, signaling cross-talk, and regulatory complexity (Fig. 1). Significantly, differences in phosphorylation-mediated signaling cascades between worms, flies and humans become apparent when examining the pathways involved in hematopoiesis and immunity. The JAK/STAT cascade, which has been implicated in hematopoiesis and cytokine signaling, is present in humans and flies. Worms, however, lack JAK kinases but do possess STAT proteins that are regulated by tyrosine phosphorylation. Like humans, flies also contain the Toll/IKK/NFκB pathway, which plays a role in the immune response to microbial organisms. No evidence of an inducible host defense system has been demonstrated in worms, consistent with the lack of this pathway in *C. elegans*. Also miss-

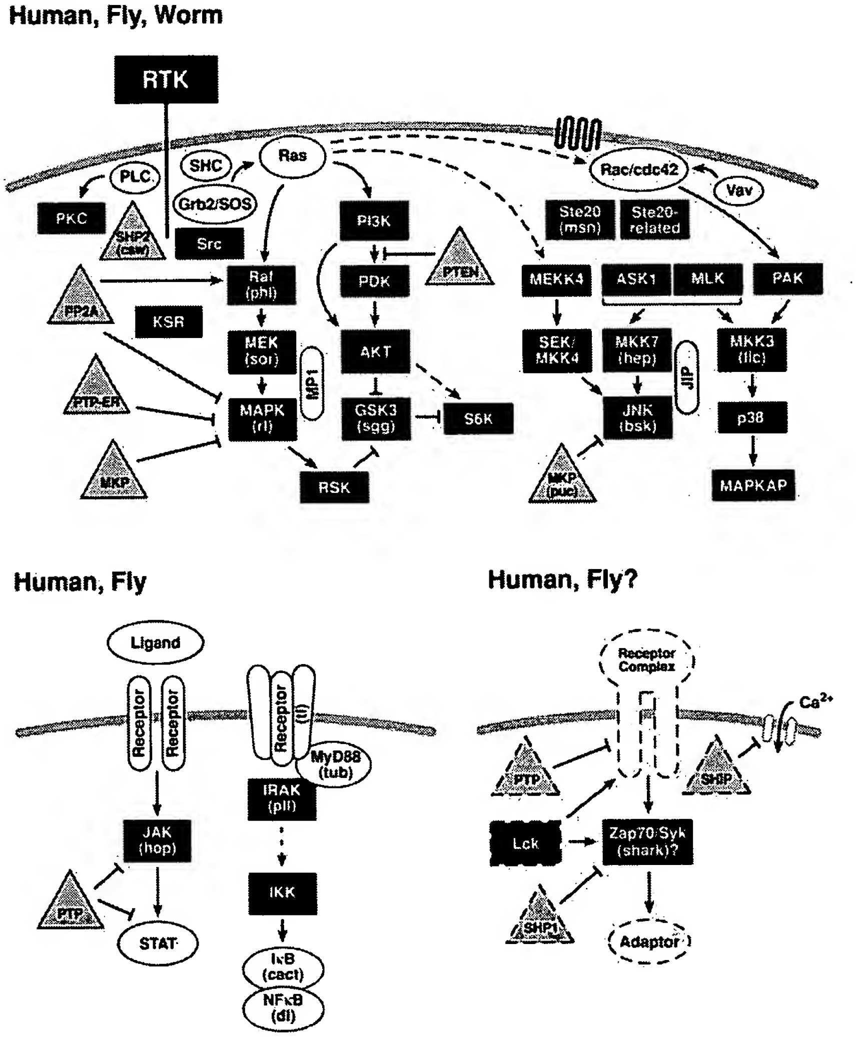


Figure 1. Comparison of the protein kinase/phosphatase signaling pathways in flies, worms, and humans (see text for description). Kinases are depicted as black rectangles, phosphatases are gray triangles, and other signaling components are in white. Shapes in dotted lines indicate mammalian proteins with no clear fly homologue; however, the function of these components in the pathway may be provided by other *Drosophila* proteins with related biochemical activities. *Drosophila* gene names are listed in parentheses.

ing in the worm are the SYK/ZAP70 kinases which play an important role in human T and B cell signaling. *Drosophila* may possess some form of this pathway as indicated by the presence of the fly SHARK kinase. The *Drosophila* SHARK kinase is a member of the SYK/ZAP70 family; however, it is most closely related to the HTK16 kinase of Hydra based on the presence of ANK repeats which are not found in any of the known mammalian SYK/ZAP70 family members (Chan et al., 1994; Ferrante et al. 1995). Exact homologues of proteins functioning with SYK/ZAP70 in the mammalian hematopoietic cascade, including the SLP-76, LAT, and BLNK adaptor proteins, the LCK and LYN kinases, and the SHP-1 and SHIP phosphatases were not revealed by the fly genome project; however, *Drosophila* proteins with related biological activities are found, namely SHP-2, inositol-1,4,5-triphosphate phosphatase, and other SRC-kinase members. Thus, further studies are required to determine whether a rudimentary form of the SYK/ZAP70 pathway does function in flies.

The completion of the *Drosophila* genome project also allows us to look globally at the pathways in which many of the newly identified fly enzymes may function. In particular, many of the proteins revealed in the *Drosophila* genome are orthologs of kinases and phosphatases known to function in the Rac/Rho/CDC42 signaling pathway (Citron, ACK2, MLK2, MEKK4, LIM-domain kinase, PAK/STE20, and DSPs members), in cell cycle regulation (CDK7, BUB1, NEK1, NEK2, CDC14, CDC7, and PP2C), and in pathways establishing asymmetry and cell polarity (LKB1, SLK1, and EMK kinases). Whether these enzymes went undetected for so many years because of functional redundancy or unappreciated phenotypes has yet to be determined.

In conclusion, ~251 protein kinases and 86 phosphatases have been identified in the *Drosophila* genome. Although the overall number of fly enzymes is lower than that found *C. elegans*, the difference is largely due to the worm-specific expansion of certain gene families. Interestingly, no large expansions or deletions of particular kinase or phosphatase gene families were uncovered by the *Drosophila* genome project. All of the previously known *Drosophila* kinases and phosphatases were detected in our analysis, confirming the relative completeness of the genome sequence data. Remarkably, almost 170 new protein kinases and phosphatases were identified by the fly genome project (Table I). The next challenge for scientists will be to determine the role of these enzymes in *Drosophila* development and physiology.

We are grateful to Celera Genomics and the Berkeley *Drosophila* Genome Project for providing access to the *Drosophila* genome sequence and predicted protein dataset. We thank Greg Plowman for his advice and insights into *Drosophila* kinase and phosphatase gene families.

Submitted: 8 June 2000
Revised: 27 June 2000
Accepted: 27 June 2000

References

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, et al. 2000. The genome sequence of *Drosophila melanogaster. Science.* 287:2185–2195.

Chan, T.A., C.A. Chu, K.A. Rauen, M. Kroiher, S.M. Tatarewicz, and R.E. Steele. 1994. Identification of a gene encoding a novel protein-tyrosine kinase containing SH2 domains and ankyrin-like repeats. *Oncogene.* 9:1253–1259.

Cheng, A., K.E. Ross, P. Kaldis, and M.J. Solomon. 1999. Dephosphorylation of cyclin-dependent kinases by type 2C protein phosphatases. *Genes Dev.* 13:2946–2957.

Chiarugi, P., P. Cirri, L. Taddei, E. Giannoni, G. Camici, G. Manao, G. Raugei, and G. Ramponi. 2000. The low M(r) protein-tyrosine phosphatase is involved in Rho-mediated cytoskeleton rearrangement after integrin and platelet-derived growth factor stimulation. *J. Biol. Chem.* 275:4640–4646.

Cohen, P.T.W. 1997. Novel protein serine/threonine phosphatases: variety is the spice of life. *Trends Biochem. Sci.* 22:245–251.

den Hertog, J. 1999. Protein-tyrosine phosphatases in development. *Mech. Dev.* 85:3–14.

Desai, C.J., Q. Sun, and K. Zinn. 1997. Tyrosine phosphorylation and axon guidance: of mice and flies. *Curr. Opin. Neurobiol.* 7:70–74.

Dick, T., S.M. Bahri, and W. Chia. 1997. *Drosophila* DPP2C1, a novel member of the protein phosphatase 2C (PP2C) family. *Gene.* 199:139–143.

Drewes, G., A. Ebneth, and E.M. Mandelkow. 1998. MAPs, MARKs, and microtubule dynamics. *Trends Biochem. Sci.* 23:307–311.

Ferrante, A.W., Jr., R. Reinke, and E.R. Stanley. 1995. Shark, a Src homology 2, ankyrin repeat, tyrosine kinase, is expressed on the apical surfaces of ectodermal epithelia. *Proc. Natl. Acad. Sci. USA.* 92:1911–1915.

Fruman, D.A., R.E. Meyers, and L.C. Cantley. 1998. Phosphoinositide kinases. *Annu. Rev. Biochem.* 67:481–507.

Gross, S.D., and R.A. Anderson. 1998. Casein kinase I: spatial organization and positioning of a multifunctional protein kinase family. *Cell Signal.* 10:699–711.

Hanks, S.K., and T. Hunter. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB (Fed. Am. Soc. Exp. Biol.) J.* 9:576–596.

Hunter, T. 1998. Anti-phosphatases take the stage. *Nat. Genet.* 18:303–305.

Laporte, J., F. Blondeau, A. Buj-Bello, D. Tentler, C. Kretz, N. Dahl, and J.L. Mandel. 1998. Characterization of the myotubularin dual specificity phosphatase gene family from yeast to human. *Hum. Mol. Genet.* 7:1703–1712.

Lowrey, P.L., K. Shimomura, M.P. Antoch, S. Yamazaki, P.D. Zemenides, M.R. Ralph, M. Menaker, and J.S. Takahashi. 2000. Positional syntenic cloning and functional characterization of the mammalian circadian mutation tau. *Science.* 288:483–492.

Maehama, T., and J.E. Dixon. 1999. PTEN: a tumour suppressor that functions as a phospholipid phosphatase. *Trends Cell Biol.* 9:125–128.

Martin-Blanco, E., A. Gampel, J. Ring, K. Virdee, N. Kirov, A.M. Tolkovsky, and A. Martinez-Arias. 1998. *puckered* encodes a phosphatase that mediates a feedback loop regulating JNK activity during dorsal closure in *Drosophila. Genes Dev.* 12:557–570.

Neel, B.G., and N.K. Tonks. 1997. Protein tyrosine phosphatases in signal transduction. *Curr. Opin. Cell Biol.* 9:193–204.

Ostanin, K., C. Pokalsky, S. Wang, and R.L. Van Etten. 1995. Cloning and characterization of a *Saccharomyces cerevisiae* gene encoding the low molecular weight protein-tyrosine phosphatase. *J. Biol. Chem.* 270:18491–18499.

Perkins, L.A., I. Larsen, and N. Perrimon. 1992. *corkscrew* encodes a putative tyrosine phosphatase that functions to transduce the terminal signal from the receptor tyrosine kinase *torso. Cell.* 70:225–236.

Peters, J.M., R.M. McKay, J.P. McKay, and J.M. Graff. 1999. Casein kinase I transduces Wnt signals. *Nature.* 401:345–350.

Plowman, G.D., S. Sudarsanam, J. Bingham, D. Whyte, and T. Hunter. 1999. The protein kinases of *Caenorhabditis elegans*: a model for signal transduction in multicellular organisms. *Proc. Natl. Acad. Sci. USA.* 96:13603–13610.

Reese, M.G., G. Hartzell, N.L. Harris, U. Ohler, J.F. Abril, and S.E. Lewis. 2000. Genome annotation assessment in *Drosophila melanogaster. Genome Res.* 10:483–501.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, et al. 2000. Comparative genomics of the eukaryotes. *Science.* 287:2204–2215.

Sells, M.A., and J. Chernoff. 1997. Emerging from the PAK: the p21-activated protein kinase family. *Trends Cell Biol.* 7:162–167.

Stein, E., A.A. Lane, D.P. Cerretti, H.O. Shloecklmann, A.D. Schroff, R.L. Van Etten, and T.O. Daniel. 1998. Eph receptors discriminate specific ligand oligomers to determine alternative signaling complexes, attachment, and assembly responses. *Genes Dev.* 12:667–678.

Whitmarsh, A.J., and R.J. Davis. 1998. Structural organization of MAP-kinase signaling modules by scaffold proteins in yeast and mammals. *Trends Biochem. Sci.* 23:481–485.

Zhang, N., J. Zhang, Y. Cheng, and K. Howard. 1996. Identification and genetic analysis of *wunen*, a gene guiding *Drosophila melanogaster* germ cell migration. *Genetics.* 143:1231–1241.

Review

The protein kinases of *Caenorhabditis elegans*: A model for signal transduction in multicellular organisms

Gregory D. Plowman*, Sucha Sudarsanam*, Jonathan Bingham*, David Whyte*, and Tony Hunter[†‡]

[†]The Salk Institute, 10010 North Torrey Pines Road, La Jolla, CA 92037; and *SUGEN, 230 East Grand Avenue, South San Francisco, CA 94080

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in April 28, 1998.

Contributed by Tony Hunter, September 7, 1999

Caenorhabditis elegans should soon be the first multicellular organism whose complete genomic sequence has been determined. This achievement provides a unique opportunity for a comprehensive assessment of the signal transduction molecules required for the existence of a multicellular animal. Although the worm *C. elegans* may not much resemble humans, the molecules that regulate signal transduction in these two organisms prove to be quite similar. We focus here on the content and diversity of protein kinases present in worms, together with an assessment of other classes of proteins that regulate protein phosphorylation. By systematic analysis of the 19,099 predicted *C. elegans* proteins, and thorough analysis of the finished and unfinished genomic sequences, we have identified 411 full length protein kinases and 21 partial kinase fragments. We also describe 82 additional proteins that are predicted to be structurally similar to conventional protein kinases even though they share minimal primary sequence identity. Finally, the richness of phosphorylation-dependent signaling pathways in worms is further supported with the identification of 185 protein phosphatases and 128 phosphoprotein-binding domains (SH2, PTB, STYX, SBF, 14-3-3, FHA, and WW) in the worm genome.

Reversible protein phosphorylation plays a central role in regulating basic functions of all eukaryotes such as DNA replication, cell cycle control, gene transcription, protein translation, and energy metabolism. Protein phosphorylation is also required for more advanced functions in higher eukaryotes such as cell, organ, and limb differentiation, cell survival, synaptic transmission, cell–substratum and cell–cell communication, and to mediate complex interactions with the external environment. Because aberrant protein phosphorylation is commonly the cause of cancer and other human diseases, a comprehensive knowledge of the key enzymes that regulate these functions can provide the basis for novel therapeutic intervention strategies.

The genomic revolution promises to provide a new paradigm for drug discovery, allowing one to selectively target the molecular basis of human disease. The completion of the *Caenorhabditis elegans* genome sequence gives us an opportunity to decipher the molecular nature of its signal transduction machinery. Several global analyses of proteins and protein domains present in *C. elegans* have been presented elsewhere (1–4), revealing that protein kinases comprise the second largest family of protein domains in worms. The three most frequently occurring protein domains found in worms are seven transmembrane chemoreceptors (650 domains, 3.5% of genome), protein kinases (496 domains, 2.6% of genome), and zinc finger C4 domains, including nuclear hormone receptors (275 domains, 1.4% of genome). A more in-depth analysis has been performed on the 535 worm proteins containing zinc-binding domains, including the C4, C2H2, and C3HC4 ring finger types (3), and on the 83 worm homeobox transcription factors (4). Here, we present a comparative analysis of the enzymes and adaptor molecules that are the key components of the protein phosphorylation signaling network present in *C. elegans*.

Identification and Classification of *C. elegans* Protein Kinases. To identify worm protein kinases, we first used an HMMER 2.1.1 (http://hmmer.wustl.edu/) profile search against the 19,099 predicted worm proteins, the finished and unfinished *C. elegans* genomic sequence, and the worm chromosome assemblies. The nucleic acid databases were first translated in all six frames, and ORFs longer than 30 amino acids were parsed into a relational database. We generated a hidden Markov model based on 70 representative yeast and human protein kinases whose catalytic domains share <50% sequence identity with each other (5). Using a similar strategy, additional profiles were generated for other protein kinase-like domains (phosphoinositide kinases, atypical A6 kinases, diacylglycerol kinases, aminoglycoside resistance kinases, and microbial kinases), protein phosphatases, and domains capable of specifically binding to phosphotyrosine (P.Tyr) or phosphoserine/threonine residues (SH2, PTB, STYX, SBF, 14-3-3, FHA, and WW domains). Scripts were written for reassembly of contiguous exons identified from genomic sequence to generate the predicted catalytic domain sequence of each kinase. Pairwise BLAST 2.0 (ftp://ncbi.nlm.nih.gov/blast/executables/) analysis was performed to identify redundant entries, and putative protein kinases with low profile scores were manually inspected to determine whether they should be included in subsequent analyses.

This analysis generated a nonredundant list of 493 protein kinase-like proteins and 21 protein kinase gene fragments from worms. This number will continue to increase as the genome is completed and the final assembly of the six worm chromosomes is achieved. Of note, we found >40 kinase domains from genomic analysis that were absent in the 19,099 worm protein dataset. These omissions result from the limitations of current protein prediction algorithms. Furthermore, numerous entries had apparent internal deletions of conserved kinase motifs, likely attributable to inappropriately assigned splice junctions. These sequences were corrected before further classification. Many of the 19,099 proteins were alternate isoforms of the same gene, in which case we included

Abbreviations: PKA, protein kinase A; MAPK, mitogen-activated protein kinase; CDK, cyclin-dependent kinase; PTK, protein-tyrosine kinase; RTK, receptor protein-tyrosine kinases; CTK, cytoplasmic protein-tyrosine kinases; STAT, signal transducer and activator of transcription; IRS, insulin receptor substrate; NLK, NEMO-like kinase; APH, aminoglycoside phosphotransferases; PTP, protein-tyrosine phosphatase.

[‡]To whom reprint requests should be addressed. E-mail: hunter@salk.edu.

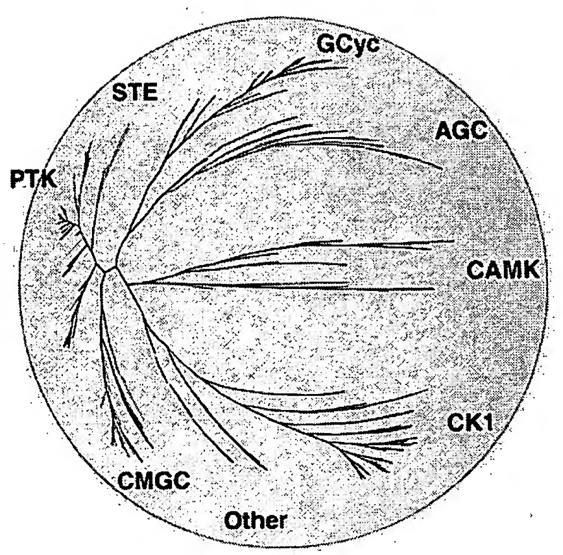


Fig. 1. Hyperbolic tree representation of *C. elegans* protein kinases. Major protein kinase groups are labeled in different colors. A JAVA tool for viewing this dendrogram can be found at www.kinase.com.

Table 1. Summary and classification of phosphoprotein signaling molecules in worms, budding yeast, and humans

Superfamily	Group	Worm	Fragments worm	Yeast	Human
Protein kinase	AGC	30	1	17	100
	CAMK	32	0	21	83
	CKI	87	7	4	5
	CMGC	42	0	24	62
	Other	62	6	29	163
	STE	28	0	15	63
	PTK	92	5	0	100
	"Worm"	27	2	0	0
	"Yeast"	0	0	4	0
	"Microbial"	7	0	6	5
	Atypical	4	0	4	11
	All	**411**	**21**	**124**	**592**
PK-like	Gcyc	26	0	1	8
	PIK	12	0	10	20
	DAGK	7	0	2	8
	YLK1	30	0	0	0
	Choline K	7	0	2	2
	All	**82**	**0**	**15**	**38**
Phosphatase	cPTP	57	4	3	25
	RPTP	26	14	0	22
	DSP	26	0	16	51
	STP	65	0	18	21
	IPP	11	0	7	7
	All	**185**	**18**	**44**	**126**
PP-Binding	SH2	73	1	1	>137
	PTB	16	0	0	>47
	STYX (DSP)	1	0	5	2
	SBF (MTM)	2	0	0	3
	14-3-3	3	0	2	>6
	FHA	11	0	14	>20
	WW	22	0	5	>32
	All	**128**	**1**	**27**	**>247**
Other	Cyclins	34	0	23	>21

only one of the proteins in our final assessment. In determining the total number of protein kinases, the three proteins determined to contain dual catalytic domains were only counted once. Many of the protein ORFs truncated the extremities of the kinase domain proteins, frequently because of their location near the end of a cosmid clone. In these cases, we searched for N- or C-terminal domains on adjacent cosmids to assist in the subsequent classification. One challenge of genomic data mining is the presence of sequence repeats. Tandem repeats and inverted repeats account for 2.7 and 3.6% of the worm genome, respectively. In addition, worms contain large regions of tandem gene duplication, ranging from hundreds of bases to >100,000 bases (1). In some cases, the genes encoded within these regions are duplicated and have nearly identical sequences. Therefore, until the chromosome sequences are fully assembled, data-mining approaches may exclude some of these duplicated genes.

A multiple sequence alignment was generated from the predicted catalytic domains of 398 of these protein kinase, which share >15% amino acid identity with other entries. The aligned proteins were then clustered by using parsimony analysis, and the results were displayed as rooted and unrooted cluster dendrograms, and as kinase "retinograms" or hyperbolic trees using a JAVA display tool (Fig. 1 and www.kinase.com*)*. The protein kinases were then classified into several kinase groups and families, based on relatedness within the kinase catalytic domain to other worm, yeast, and vertebrate protein kinases. Further classification was performed by searching for noncatalytic domains linked to the kinase domain, including predicted transmembrane regions, SH2 domains and SH3 domains, and Ig and fibronectin Type III domains.

Table 1 presents a summary of our classification of the 411 protein kinases and 82 protein kinase-like motifs. A more detailed table of these proteins, along with basic informatics tools for retrieval and alignment of these sequences can be found on our web site at www.kinase.com. Table 1 also summarizes the results of a similar analysis of the completed yeast genome and of an ongoing effort from publicly available human expressed sequence tag and genomic databases. From this classification, we can now determine which protein kinases are conserved between yeast and worms, we can speculate on the origin of the protein kinase superfamily, and we can identify kinases that are yeast-specific and those that are restricted to higher eukaryotes. We tentatively identify "worm-specific" protein kinases, based on their absence from current mammalian expressed sequence tag and nucleic acid databases. However, a final assessment will have to await completion of the *Drosophila* and human genome sequences. We also elaborate on some of the protein kinases and signaling pathways that evolutionarily appear only in more complex organisms such as vertebrates.

In this review, we use the term "orthologues" to refer to proteins of different species that are believed to have a common ancestor and have an evolutionarily conserved function. Orthologous proteins typically have similar domain structure and share extended sequence similarity outside of their catalytic domains. Homologous proteins also share extended sequence similarity, but to a lesser degree than orthologues, and are not expected to complement one another functionally. However, within large protein superfamilies such as protein kinases, G protein coupled receptors, and nuclear hormone receptors, there is not a single expectation value that can be used to categorize all members definitively, and final classification will require experimental validation.

Yeast- and Fungal-Specific Kinases. The first complete eukaryote sequence, that of the budding yeast *Saccharomyces cerevisiae*, was reported in 1996 (6). Shortly thereafter, we presented a comprehensive analysis and classification of yeast protein kinases (7). Now, with the availability of a second eukaryotic genome, *C. elegans*, we can perform a similar analysis and make more informed generalizations on which of these protein kinases are unique to yeast or fungi, and also on which protein kinases evolved during the emergence of multicellular organisms and are therefore not represented in yeast or fungi.

We now identify a total of 24 yeast-specific protein kinases and an additional 3 that are currently restricted to yeast and worms. Originally we defined four protein kinase subfamilies, containing a total of 18 members, to be yeast specific [protein kinase A (PKA)-related, RAN, ELM, and NPR/HAL5 families]. These remain yeast- or fungal-specific, as no close homologues are present in worms, and none have yet been described in vertebrates. However, the ELM family could be considered as a subfamily of the CAMK group. Rim15 is a yeast-specific kinase that is related to *Schizosaccharomyces pombe* Cek1, and its similarity to budding yeast YNL161w places it as a distant member of the NDR family kinases. Two other protein kinase subfamilies, containing a total of five members, were originally recognized as having only distant homologues in higher organisms (NEK-like and PIM-like families). The prototype of the NEK-like family, YNL020C, has a homologue in worms, but not in mammals, although its C-terminal tail has a predicted coiled-coil structure related to numerous mammalian protein kinases (e.g., SLK/PLKK, TAK1). The two yeast PIM-like family members have catalytic domains related to worm and mammalian protein kinases, but have a unique N-terminal domain.

Members of the NPR/HAL5 family are involved in ion homeostasis, polyamine transport, nutrient uptake, and response to nitrogen starvation, whereas Elm1 initiates a protein kinase cascade controlling pseudohyphal growth (8). Members of the RAN family are related to fission yeast Ran1/Pat1, which regulates the switch between vegetative growth and meiosis. Because these are fungal-specific responses, it is not surprising that these protein kinases are restricted to lower eukaryotes.

A second set of "unique" yeast protein kinases was originally defined because they had no close homologues in other species (7). Most of these yeast protein kinases now have both worm and vertebrate orthologues (Cdc5, Ipl1, Ire1, Vps15, YGL180W/Apg1, Swe1, Spk1, Gcn2, YBR274W, YGR262C, and Bub1). Exceptions among this list of unique yeast protein kinases are YPL236C and Mps1, which have orthologues in humans, but not in worms; YKL116C, which is distantly related to the EMK-family, yet has only weak homologues in worms and humans; and YKL171W, YGR052W, and YPR106W, which remain yeast specific protein kinases. Two sequences that were excluded from our previous analysis of yeast protein kinases deserve mention. The budding yeast protein Iks1 can be classified as a yeast-specific protein kinase because it still has no homologues in worms or other species whereas another yeast kinase-like sequence, SCY1, has orthologues in *C. elegans* and *Arabidopsis*, but none thus far in vertebrates. A *S. pombe* protein, which is distantly related to SCY1, also has a single worm orthologue.

Worm-Specific Protein Kinases. Which protein kinases are specific to worms? Protein kinases that are absent from yeast yet present in worms are likely to be involved in the complex signal transduction pathways that are required for the existence of multicellular organisms. These might include protein kinases involved in cell–substratum and cell–cell adhesion, transmembrane signaling in response to humoral factors, protein kinases involved in cell survival or programmed cell death, and protein kinases whose signals regulate metazoan-specific transcription factors, particularly those containing Zn-finger domains.

In the absence of complete genome sequences of other multicellular eukaryotes, we tentatively classify 165 protein kinases (plus 9 protein kinase fragments) as worm-specific. The majority (134, 80%) fall into three groups (CK1, FER, and KIN-15) whereas the others are distant members of common protein kinase families or belong to worm specific subfamilies. Five protein kinase subfamilies, containing a total of 12 members, can tentatively be defined as worm-specific (C04G2.10, K08B4.5, K09C6.7, R107.4, and ZK177.2-families). An additional 15 unique worm protein kinases are also identified, which to date have no close homologues in yeast, worms, or in higher organisms. However, mammalian homologues of some of these worm protein kinases are already beginning to appear in publicly available expressed sequence tag databases, and assignment of a protein kinase as being truly worm-specific will have to await the completion of the *Drosophila* and human genome sequences.

Members of four other protein kinase or kinase-like subfamilies are disproportionately represented in worms compared with humans. Clusters of 5–9 members of each of these families are localized to short regions (<1 megabase) of chromosomes II and IV, suggesting they may each have expanded as a result of extensive tandem gene duplication. The chromosomal density of protein kinases is graphically depicted on our web site at www.kinase.com. The four gene families are the CK1-family, the KIN-15-family of receptor protein-tyrosine kinases, the FER-family of cytoplasmic protein-tyrosine kinases, and the kinase-like domains of the receptor guanylyl cyclases.

CK1 family. The worm genome contains 87 CK1 (casein kinase I) members (plus 7 additional partial catalytic domains) whereas there are only 4 known members in budding yeast and 6 in humans. Genetic evidence from the yeast homologues suggests CK1s may be involved in DNA repair and cell division, and mammalian CK1s have been shown to phosphorylate p53 in G1 and G2, possibly affecting cell sensitivity to DNA damage at these checkpoints (9). Little is known regarding the function of CK1s in worms, but the enormous arborization and diversification of this kinase family may be an adaptation allowing for enhanced DNA repair in response to excessive exposure to environmental mutagens.

KIN-15/16 family. C. elegans contains 16 members of a unique family of receptor protein-tyrosine kinases whose presence to date is restricted to this species. These transmembrane proteins have unusually short (<50-aa) extracellular domains, and many are clustered within the genome, as though they arose through tandem gene duplication. The prototype members of this family, KIN-15 and KIN-16, are expressed in the hypodermal syncytium, which expands by cell fusion during larval development (10). Compared with wild-type worms, KIN-15 and KIN-16 deletion mutants produce fewer embryos and rarely develop into adults, but, when they do mature, they typically exhibit extrusion of the gonads through the vulva (11). Therefore, KIN-15/16 appear to be essential genes, yet may undergo variable compensation by 1 of the 14 other homologues. One of the KIN-15 clusters is interspersed with chitinase genes, which are known to function in cell wall morphogenesis during the molting process and in fungal resistance. Expansion of this region may have been necessary during evolution to facilitate this aspect of larval development. An alternative function for KIN-15-family kinases is suggested by the fact that overexpression TKR-1 (C08H9.5) causes a 40–100% extension of life expectancy in worms (12). Unlike other life extension (*age*) mutants, TKR-1 transgenics do not form dauers, and their longevity has been attributed to an increased resistance to ultraviolet and thermal stress.

FER family. The worm genome contains 42 members (plus 2 additional partial catalytic domains) of the FER-family of single SH2-containing cytoplasmic protein-tyrosine kinases. Most of these genes are interspersed throughout the worm genome; however, nine members reside within a 1.1-megabase region on chromosome IV. Unfortunately, no literature is available on the function of any of these protein kinases in worms, but the two mammalian homologues, FER and FES, have been demonstrated to play a role in cell adhesion, to signal downstream of cytokine receptors, and to function as oncogenes (13). Conceivably, additional human representatives will be revealed on completion of the human genome sequence, possibly with restricted expression. Alternatively, their function may be replaced in humans by expansion and diversification of non-FER cytoplasmic protein-tyrosine kinases, of which worms have only 10 whereas humans have at least 34. Most evident is a dramatic expansion of SRC-family kinases and emergence of ZAP70 and JAK family kinases in higher eukaryotes that are not found in the worm genome.

Conserved Metazoan Protein Kinase Signaling Transduction Pathways. Worms provide an elegant model system for studying signal transduction. This transparent animal is comprised of 959 somatic cells plus 131 cells destined for programmed cell death. The *C. elegans* hermaphrodite contains 302 neurons and 81 muscle cells and has a brain, reproductive system, and digestive tract (ref. 14; http://dauerdigs.biosci.missouri.edu/Dauer-World/Wormintro.html). It provides a complex yet tractable system for studying development, metabolism, aging, and behavioral responses to a number of stimuli. Regulation of many of these processes is carried out through signal transduction pathways that are also present in humans. Not surprisingly, all of the major protein kinase groups found in worms are also conserved in humans (15). The number of protein kinases classified into each major group from yeast and worms, along with a current estimate from humans, is provided in Table 1. These numbers represent a current analysis, but new protein kinases are being discovered every month as the worm genome sequencing project continues. Some of these entries may also represent pseudogenes containing frameshifts that result in incomplete translation into a full kinase catalytic domain.

AGC Group. The AGC group of worm protein kinases contains representatives of many of the known types of cyclic nucleotide-dependent, NDR or DBF2, and ribosomal S6 kinase families. Worms also contain members of the cGMP-dependent kinase (PKG), RSK, and G-protein coupled receptor kinase families that are absent from budding yeast. Two of the S6 kinase members have dual catalytic domains similar to vertebrate RSK enzymes, where the N-terminal domain clusters into the AGC group and the C-terminal kinase domain is most related to the CaMK group. Worms have four members of the AKT family, two being close orthologues of mammalian AKT1/PKB/RACα, and two related to the AKT upstream kinase, PDK1. AKT is a mammalian protooncoprotein regulated by phosphatidylinositol 3-kinase (PI3-K), which appears to function as a cell survival signal to protect cells from apoptosis (16). Insulin receptor, RAS, PI3-K, and PDK1 all act as upstream activators of AKT whereas the lipid phosphatase PTEN functions as a negative regulator of the PI3-K/AKT pathway (17). Downstream targets for AKT-mediated cell survival include the proapoptotic factors BAD and Caspase9 and transcription factors in the forkhead family, such as DAF-16 in the worm. AKT is also an essential mediator in insulin signaling, in part because of its use of GSK-3 as another downstream target. Each of these components of the AKT/PI3-K pathway is conserved in worms, providing a powerful system for genetic dissection of a major cell survival signal.

The cAMP-dependent protein kinases (PKA) consist of heterotetramers comprised of two catalytic (C) and two regulatory (R) subunits, in which the R subunits bind to the second messenger cAMP, leading to dissociation of the active C subunits from the complex. Worms have two PKA catalytic domains and two regulatory subunit genes (R07E4.6 and ZK370.4). Additional cNMP-binding domains are present in the two worm representatives of the PKG family, in several cNMP-gated ion channels, and in a cAMP-regulated guanine nucleotide exchange factor (T20G5.5).

CaMK Group. In the CaMK group, the most abundant representatives include Ca^{2+}/calmodulin-regulated and AMP-dependent protein kinases and EMK-related kinases. Worms also contain members of the death-associated protein kinase, mitogen-activated protein kinase (MAPK)-associated protein kinase, myosin light chain kinase, and phosphorylase kinase families that are absent from budding yeast. All of these protein kinase families have likely evolved as a result of the demands of multicellularity and the emergence of complex organ systems. For example, even though yeast have myosin homologues, they lack myosin light chain kinases. These protein kinases have presumably evolved to regulate myosin during muscle contraction. A worm contig still under construction appears to contain a phosphorylase kinase catalytic γ subunit orthologue, consistent with the presence of two orthologues of the noncatalytic phosphorylase kinase α subunits, which facilitate calmodulin-binding and are required for activation of the mammalian holoenzyme.

Worms lack a homologue of the mammalian Trio-family kinases. Trio is a large multidomain protein kinase containing Ras and Rho guanine exchange factor domains in addition to PH, SH3, and spectrin domains (18). Trio may link Rho and Rac signaling pathways and appears to be involved in the cytoskeletal changes required for cell migration. Although worms lack a member of this kinase family, they do have at least two proteins related to the entire noncatalytic domain of Trio (UNC-73 and F55C7.7).

We have also identified a forkhead homology (FHA) domain-containing CHK2 orthologue in worms. In yeast, Spk1/Rad3 functions as a DNA damage checkpoint sensor through its FHA domain interacting with phosphorylated Rad9 (19). Although no close orthologue of Spk1 exists in metazoans, this function appears to be replaced by CHK2/CDS1, which is phosphorylated in response to DNA damage and may work in conjunction with CHK1 kinase to phosphorylate CDC25C to prevent premature entry into mitosis (20).

CMGC Group. In the CMGC group of serine/threonine kinases, all of the main subfamilies are conserved between yeast, worms, and mammals, including cyclin-dependent kinase (CDK), MAPK, GSK-3, and CLK. An exception is the RCK family, which is absent from yeast but has two members in worms and at least seven in humans. The worm RCK kinases are most similar to mammalian MAK, or male germ cell-associated kinase, which has been implicated in spermatogenic meiosis and in signal transduction pathways for sight and smell. Worms have 14 CDKs (compared with 5 CDKs in yeast) including orthologues of CDC2, CDK3, CDK5, CDK7, and CDK8, and contain 34 cyclins, compared with 23 in budding yeast (Table 1), including one cyclin H orthologue, which we predict will interact with worm CDK-7 to generate a functional cyclin-activated kinase.

Worms have 14 MAPKs, compared with 6 in yeast and at least 14 in humans. The worm MAPKs include representatives of each of the major types of MAPKs: ERK/MAPK, JNK/SAPK1, p38/SAPK2, BMK/ERK5, and NEMO-like kinase (NLK) (21). In budding yeast, three protein kinase families (the prototypes being Ste20, Ste11, and Ste7) function upstream of the MAPKs to generate at least four distinct MAPK signaling pathways that mediate the response to pheromone, nutritional starvation, and cellular or osmotic stress. In multicellular organisms, these MAPK cascades have evolved to mediate responses to diverse signals including growth factors, mitogens, hormones, and cytokines, in addition to the more primitive stress responses to anoxia, heat shock, and osmotic stress.

STE Group: MAPK Pathways. The STE family refers to the three classes of protein kinases that lie sequentially upstream of the MAPKs. In worms, this group includes 10 STE7 (MEK or MAPKK) kinases, 2 STE11 (MEKK or MAPKKK) kinases, and 12 STE20 (MEKKK) kinases. Based on the number of MAPK and STE-family kinases in *C. elegans*, we predict worms will contain at least 8–10 MAPK pathways. In humans, several protein kinase families that bear only distant homology with the STE11 family also operate at the level of MAPKKKs, including RAF, MLK, TAK1, and COT. Except for COT, worms also have orthologues of each of these kinases. Because crosstalk takes place between protein

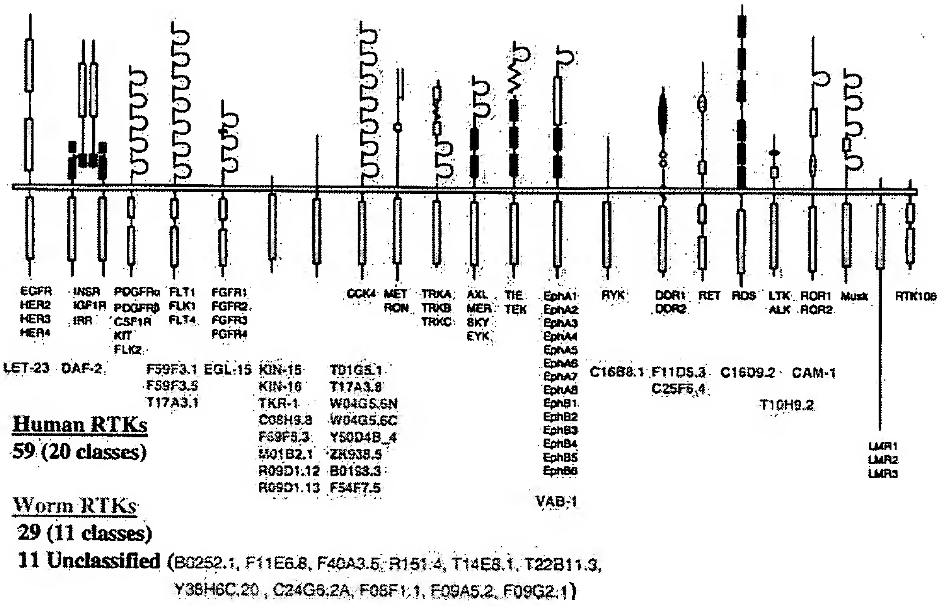


Fig. 2. Schematic representation of the human and C. *elegans* receptor protein-tyrosine kinase families. Catalytic domains are shown in yellow. The names of the human RTKs are in black, and the names of the worm RTKs are in red.

kinases functioning at different levels of the MAPK cascade, the large number of STE family kinases could translate into an enormous potential for upstream signal specificity and diversity.

Protein-Tyrosine Kinase Group: Receptor Protein-Tyrosine Kinases (RTKs). The largest group of protein kinases in worms are the protein-tyrosine kinases (PTKs), with 92 members and 5 fragments. We predict this will also remain the largest group of protein kinases in higher eukaryotes, including humans, where the current count is ≈100. These numbers are impressive when one considers that this family is absent from budding yeast. Yeast, however, do have a "budding" tyrosine phosphorylation signaling system, with several dual-specificity kinases (CLK-like, Ste7/MEK family, Swe1, Spk1/Rad53, Mps1), an atypical A6 PTK, 3 protein-tyrosine phosphatases, 16 dual-specificity and low molecular weight phosphatases, and 6 "infant" P.Tyr-binding proteins comprising an apparently nonfunctional SH2 domain protein and 5 phosphatase-like STYX domains. Budding yeast lack PTB domains, and none of the six potential P.Tyr-binding domains have been functionally verified.

The 92 worm PTKs can be further classified into receptor protein-tyrosine kinases (RTKs) and cytoplasmic protein-tyrosine kinases (CTKs) based on the presence or absence of a transmembrane domain and SH2 or SH3 domains. Based on this analysis, the worm genome contains 40 RTKs and 52 CTKs. The 40 RTKs include 16 members of the worm-specific KIN-15-family, 13 RTKs with orthologues representing 10 of the 20 families of human RTKs, and 11 RTKs that remain unclassified with no identifiable mammalian counterpart (Fig. 2). Genetic studies in worms support the classification of five of these worm–human pairs, including LET-23/EGF receptor, DAF-2/insulin receptor, EGL-15/FGF receptor, CAM-1/ROR1 receptor, and VAB-1/EPH receptor, and each of these orthologous pairs mediates similar functions in worms and man, with specificity for epidermal, metabolic, mesodermal, and neuronal signaling pathways.

Based on extracellular domain homologies, we also predict three worm orthologues of PDGFR/FLK/VEGFR, two for DDR, and one each of RYK, ROS, and LTK/ALK. Two of the unclassified RTKs have weak similarity to MET, but not enough to warrant inclusion into this family. Missing in *C. elegans* are TRK/nerve growth factor receptors, AXL/TYRO3, TIE/angiopoietin receptor, RET/GDNF receptor, and MUSK family members. Identification of three members of the PDGFR/VEGFR family is significant, as they emerged only through analysis of the genomic data and failed to be properly identified from a recent analysis of the predicted 19,099 proteins. Each of these receptors contains multiple Ig-like extracellular domains and a single split kinase domain with closest homology to human FLT1/VEGFR1 and the *C. elegans* KIN-15 family. However, they are likely to represent early ancestors to both the FLK and PDGFR kinase lineages. Expression of the mammalian FLK/VEGFR RTKs is primarily restricted to endothelial cells, and they play important roles in the early differentiation of hematopoietic and endothelial lineages as well as in normal and pathologic angiogenesis in the adult. However, because worms lack a vasculature, the function of these receptors is not obvious. The formation of mammalian vasculature is reminiscent of the process by which networks of branching tubes develop into the lung and kidneys. Invertebrate VEGFRs may therefore be involved in processes that later evolved into a program for limb and organ development in vertebrates.

Surprisingly little is known about how the ligand-activated VEGFRs mediate these effects. Gene knockout studies in mice suggest that A-RAF or MEKK1 may function downstream of VEGFRs, and recent evidence implicates the involvement of STATs (signal transducer and activator of transcription) in VEGFR signaling (22). Genomic analysis reveals two worm orthologues of STATs (Y51H4, Y43D4 unfinished and F58E6.1), making the VEGFR-STAT association an attractive area for further investigation. STATs contain an SH2 domain, a tyrosine phosphorylation domain, and a DNA-binding domain, and function in a unique JAK-STAT signaling pathway. Extensive studies in mammalian systems have established a model in which JAK kinases are constitutively bound to the cytoplasmic portion of cytokine receptors and are activated on receptor dimerization, facilitating recruitment of STATs to the receptor complex. Subsequent STAT phosphorylation leads to their dimerization and translocation to the nucleus, where they function directly as transcription factors. *Drosophila* and *Dictyostelium* STATs both regulate cell division and pattern formation (23, 24). *Drosophila* STAT has been genetically and biochemically linked to a JAK-STAT signal transduction pathway that regulates pair-rule genes and hematopoiesis. *Dictyo-*

stelium STAT plays an essential role during the differentiation and aggregation of independent spore cells into stalk cells in response to the chemical signal referred to as differentiation-inducing factor. Furthermore, the *Dictyostelium* AX2 PTK has a second kinase-like domain found only in JAK-family kinases, suggesting the existence of a signaling network similar to that in flies and mammals. However, worms have no cytokines, no cytokine receptors, and no JAK-family kinases. Possibly, the JAK kinase function is replaced by a worm-specific FER kinase, or the STATs may have initially evolved to serve an alternative purpose. Mammalian STATs are also involved in signaling through receptors for growth factors such as EGF, PDGF, VEGF, and angiopoietins. Because the EGF and VEGF signaling systems are present in worms, it is tempting to speculate that these represent the primordial raison d'être for STATs.

In general, related RTKs bind related ligands. In humans, there are at least 12 ligands, encoded by 10 genes, that have been shown to bind selectively to at least one of the four known EGFR-family members. Each of these ligands shares a conserved six-cysteine pattern in its receptor binding domain. In worms, LIN-3 has been shown to function as a LET-23 ligand. Although EGF motifs are prevalent in worms, we have identified three EGF-like proteins (F58G4.4, Y69H2.2, YG70G10A.2) that, in addition to the six cysteines, conserve many of the crucial receptor-binding residues and are juxtaposed next to a putative transmembrane domain, in a pattern similar to all known EGFR-family ligands. Worms also contain at least 3 FGF-like ligands, 12 insulin-like ligands (many more on inclusion of relaxin-related ligands), 2 distant homologues of VEGF, and 4 ephrin-related ligands, some of which would be predicted to bind to their cognate receptors.

Orthologues of other RTK ligands prove more difficult to identify empirically. We see no evidence for a bona fide PDGF or NGF, and searches for ligands for MET, TIE, and AXL-family RTKs are confounded by their similarity to plasminogen, fibrinogen, and fibrillin, respectively. Furthermore, except for weak homologues of MET, these three RTK families are absent from worms. Nevertheless, the significance of a putative Ang2-like protein (Y43C5A.2) in the absence of a TIE-family RTK remains to be determined.

Protein-Tyrosine Kinase Group: CTKs. Most of the 52 CTKs in worms belong to the single SH2-containing FER family. Of the remaining 10 CTKs, there are 2 orthologues of the SH3-containing ACK, and 1 each of FYN (SRC family CTK), FRK, CSK, ABL, and FAK, plus 3 unclassified CTKs. In vertebrates, CSK negatively regulates FYN-family kinases by phosphorylation of a C-terminal tyrosine facilitating a conformational change through an intramolecular SH2-P.Tyr interaction (25). We predict a similar functional interaction between worm FYN and CSK. Co-evolution of this regulatory pair suggests even early metazoans required a means to dampen signaling through CTKs. Notably absent in worms are protein kinases related to the ZAP70 and JAK CTKs, whose primary role in mammals is in signaling through the T cell and cytokine receptors, both of which represent more specialized pathways not present in worms. Humans have eight SRC-family kinases whereas worms have only one. This redundancy has confounded efforts to dissect out the precise role of these CTKs in human biology, often requiring "triple knockouts" to demonstrate a deficiency. The simplicity of non-FER-like CTKs in worms may be helpful in placing these CTKs within specific signaling cascades.

Protein-Tyrosine Kinases: Adaptor and Docking Molecules. Ligand activation of RTKs results in tyrosine phosphorylation of both the receptor itself (autophosphorylation) and of downstream substrates. These phosphorylated tyrosines then function as attachment sites for proteins containing SH2 and other P.Tyr-binding domains. We have identified 74 proteins containing a total of 77 SH2 domains in worms. The majority of these SH2 domains are in CTKs, two are present in a SHP2-related PTP, and the remainder are predicted to represent orthologues of a variety of adaptor molecules, including phospholipase Cγ, CBL, CIS4/SOCS5, CRK, NCK, SEM-5/GRB2, SHC, tensin, STAT, and VAV. Worms also contain at least 16 PTB domains, which in some cases have been found to interact specifically with tyrosine phosphorylated proteins. Worm PTB-containing proteins include orthologues of SHC, which also contains an SH2 domain, neuronal transmembrane protein X11, and an insulin receptor substrate (IRS) family member. The mammalian X11 PTB domain does not to bind to P.Tyr, so we anticipate only a few of these worm domains will function as P.Tyr-binding domains. Additional potential phosphoprotein-binding domains identified in worms include three 14-3-3 domains, 22 WW domains, and 11 FHA domains.

IRS-1 and IRS-2 are major substrates of the insulin receptor RTK in mammals, and disruption of IRS-2 in mice leads to metabolic defects similar to diabetes. Worms have multiple insulin-like peptides, a receptor, and an IRS orthologue, demonstrating the early origins of metabolic regulation in multicellular organisms. The presence of such a diverse array of adaptor molecules underscores the utility of worms as a model for understanding mammalian signal transduction.

Other Protein Kinases. Approximately 15% of the worm protein kinases do not fall into one of the six major groups but include smaller families with representatives in higher eukaryotes, including CHK1, DYRK, MLK, TAK1, PIM, RAF, STKR, and the mitotic kinases (BUB1, AURORA, PLK, and NIMA/NEK). Recent genetic and biochemical data place TAK1 (transforming growth factor β-associated kinase) on a MAPK-like pathway at the level of a MAPKKK acting upstream of the MAPK-family member NLK. The worm orthologues of TAK1 and NLK regulate Wnt-mediated cell polarization during embryogenesis (21). Biochemical data also demonstrate that this MAPK-like pathway negatively regulates Wnt signaling because NLK phosphorylates the TCF/LEF HMG transcription factors, thereby inhibiting Wnt-regulated binding of the β-catenin-TCF complex to DNA. Both of these pathways are conserved between mammals and worms. The likely orthologous human/worm pairs on the TAK1 MAPK-like pathway include TAK1/MOM-4, NLK/LIT-1, and TCF4/POP-1. Upstream regulators may include TGFβ1/DBL-1, TGFβ type I receptor/SMA-6, TGFβ type II receptor/DAF-4 (worms have three receptor serine kinases). Additional components of the Wnt-signaling pathway, such as cadherin, the adenomatous polyposis coli tumor suppressor gene (APC), disheveled, and GSK-3 kinase are also present in worms, suggesting that there may be a primordial connection between polarized control of cell division/migration and cellular transformation in vertebrates (26).

Microbial-Like Kinases: Origin of Protein Kinases? The availability of the sequence of the first complete metazoan genome, combined with the sequence of budding yeast and several prokaryotic and *Archaea* genomes, provides an excellent opportunity to reassess current theories on the evolutionary origin of protein kinases. Pkn1 is a bacterial protein kinase-like sequence first described in the Gram-negative bacteria *Myxococcus xanthus*, which functions in growth and differentiation and in the ability of this prokaryote to form a fruiting body in response to nutrient starvation. Pkn-related proteins are present in other prokaryotes, including *Streptomyces*, *Bacillus*, *Mycobacterium*, *Pseudomonas*, *Chlamydia*, and *Synechocystis*, where they are involved in virulence, secondary metabolism, sporulation, and complex growth cycles (27). However, there are no Pkn homologues in bacteria with less complex life cycles, such as *Escherichia coli*, and *Haemophilus influenzae*, or in any *Archaea*, suggesting they may have been acquired by horizontal transmission from an early eukaryote, and are unlikely to represent the ancestral founders to protein kinases.

In our kinase profile searches of the worm genome, we detected several entries with low profile scores, yet with significant (E value $< 10^{-2}$) random expectation (E) values. Most of these contained similarity to kinase subdomains I, II, and VI, containing the consensus GxGxxGxV, VAVK, and HxDxxxxN motifs, respectively. Upon further analysis, many of these entries could be classified into distinct families designated ABC1, RI01, YGR262, diacylglycerol kinase, choline/ethanolamine kinases, and the YLK1-antibiotic resistance kinases. The first three families are named after their prototypic members in *S. cerevisiae* (27).

Worms contain three proteins related to the budding yeast ABC1. The yeast protein is required for the assembly of the mitochondrial cytochrome *c* reductase complex, which functions as an electron carrier during oxidative phosphorylation to generate ATP (28). ABC1 homologues are present in numerous prokaryotes, including *Mycobacterium*, *Clostridium*, *Rickettsia*, *Synechocystis*, *Azobacter*, and *Enterobacteriaceae* such as *E. coli* and *Providencia stuartii*, in addition to the *Archaea*, *Methanobacterium*. ABC1-like proteins are also present in eukaryotes, including fission yeast, *Arabidopsis*, worms, and mammals. Although ABC1 homologues are absent from bacteria such as *Mycoplasma*, *Bacillus*, *Haemophilus*, *Helicobacter*, and spirochetes, their presence in other prokaryotes, *Archaea*, and eukaryotes positions them as likely representatives of the primordial protein kinase, which was the progenitor of the eukaryotic protein kinase family. Based on their recognized role in mitochondrial ATP production and because they maintain many of the structurally important residues and motifs involved in ATP binding, the ABC1-family proteins may either bind ATP or act as phosphotransferases. Conceivably, the ABC1 proteins transfer phosphate to proteins as part of a feedback loop to sense mitochondrial ATP levels.

The RI01 family has three representatives in worms and is named after one of the two homologues in budding yeast. There are also representatives from several *Archaea* species, but none from bacteria, making them a less attractive candidate as a progenitor to the protein kinase lineage.

Atypical Protein Kinases and Protein Kinase-Like Domains. Worms contain 26 kinase-like domains present in receptor guanylyl cyclases (there are 10 additional soluble guanylyl cyclases), and at least 7 diacylglycerol kinases, 7 choline/ethanolamine kinases, and 30 YLK1-related antibiotic resistance kinases. Each of these families contain short motifs that were recognized by our profile searches with low scoring E-values, but *a priori* would not be expected to function as protein kinases. Instead, the similarity could simply reflect the modular nature of protein evolution and the primal role of ATP binding in diverse phosphotransfer enzymes. However, two recent papers on a bacterial homologue of the YLK1 family suggests that the aminoglycoside phosphotransferases (APHs) are structurally and functionally related to protein kinases (28, 29). There are over 40 APHs identified from bacteria that are resistant to aminoglycosides such as kanamycin, gentamycin, or amikacin. The crystal structure of one well characterized APH reveals that it shares >40% structural identity with the two-lobed structure of the catalytic domain of cAMP-dependent protein kinase (PKA), including an N-terminal lobe composed of a five-stranded antiparallel β sheet and the core of the C-terminal lobe, including several invariant segments found in all protein kinases (29). APHs lack the GxGxxG normally present in the loop between β strands 1 and 2 but contain 7 of the 12 strictly conserved residues present in most protein kinases, including the HGDxxxN signature sequence in kinase subdomain VIB (29). Furthermore, Daigle *et al.* (30) have demonstrated that this APH also exhibits protein-serine/threonine kinase activity, suggesting that the worm YLK1-related molecules may indeed be functional protein kinases.

The eukaryotic lipid kinases (PI3Ks, PI4Ks, and PIPKs) also contain several short motifs similar to protein kinases but otherwise share minimal primary sequence similarity. However, once again, structural analysis of PIPKIIβ defines a conserved ATP-binding core that is strikingly similar to conventional protein kinases (31). Three residues are conserved among all of these enzymes, including (relative to the PKA sequence) Lys-72, which binds the α-phosphate of ATP, Asp-166, which is part of the HRDLK motif, and Asp-184, from the conserved Mg^{2+} or Mn^{2+} binding DFG motif (31). The worm genome contains 12 phosphatidylinositol kinases, including 3 PI3-kinases, 2 PI4-kinases, 3 PIP5-kinases, and 4 PI3-kinase-related kinases. The latter group has four mammalian members (DNA-PK, FRAP/TOR, ATM, and ATR), which have been shown to participate in the maintenance of genomic integrity in response to DNA damage and exhibit true protein kinase activity, raising the possibility that other PI-kinases may also act as protein kinases. Regardless of whether they have true protein kinase activity, PI3-kinases are tightly linked to protein kinase signaling, as evidenced by their involvement downstream of many growth factor receptors and as upstream activators of the cell survival response mediated by the AKT protein kinase.

There are several proteins with protein kinase activity that appear structurally unrelated to the eukaryotic protein kinases. These include *Dictyostelium* myosin heavy chain kinase A, *Physarum polycephalum* actin-fragmin kinase, the human A6 PTK, human BCR, mitochondrial pyruvate dehydrogenase and branched chain fatty acid dehydrogenase kinase, and the prokaryotic "histidine" protein kinase family. Worms lack representatives of the actin-fragmin kinases, BCR, and bacterial histidine kinases yet do contain a single representative of the other classes of atypical kinases and two homologues of the A6-related PTKs. The single worm orthologue of the *Dictyostelium* myosin heavy chain kinase A (32) proves to be the worm eukaryotic elongation factor 2 kinase (33). The slime mold, worm, and human eukaryotic elongation factor 2 kinase homologues have all been demonstrated to have protein kinase activity, yet they bear little resemblance to conventional protein kinases except for the presence of a putative GxGxxG ATP-binding motif (33).

The so-called histidine kinases are abundant in prokaryotes, with >20 representatives in *E. coli*, and have also been identified in yeast, molds, and plants. In response to external stimuli, these kinases act as part of two-component systems to regulate DNA replication, cell division, and differentiation through phosphorylation of an aspartate in the target protein (34). To date, no "histidine" kinases have been identified in metazoans, although mitochondrial pyruvate dehydrogenase (PDK) and branched chain α-ketoacid dehydrogenase kinase are related in sequence. PDK and branched chain α-ketoacid dehydrogenase kinase represent a unique family of atypical protein kinases involved in regulation of glycolysis, the citric acid cycle, and protein synthesis during protein malnutrition. Structurally, they conserve only the C-terminal portion of "histidine" kinases, including the G box regions. Branched chain α-ketoacid dehydrogenase kinase phosphorylates the E1α subunit of the branched chain α-ketoacid dehydrogenase complex on Ser-293, proving it to be a functional protein kinase (35). Although no bona fide "histidine" kinase has yet been identified in worms or humans, they do contain PDK homologues (one in worms and five in humans). However, the paucity of PDKs in worms makes it unlikely that they fill in for the absence of "histidine" kinases in metazoans. Instead, these signaling cascades have more likely been replaced by pathways initiated through RTKs.

Based on these examples of atypical protein kinases present in the worm genome, we predict additional worm protein kinases will be functionally identified that lack any of the obvious motifs conserved in the conventional members. Indeed, various biochemical data point to the existence of true histidine, lysine, and arginine kinases in metazoans, yet their structural identity remains a mystery.

Protein Phosphatases. Because of their important role in signal transduction, it is not surprising that the activity of protein kinases must be tightly regulated. This is accomplished through autoinhibition, autophosphorylation, transphosphorylation, dimerization, and cellular localization. Equally important is the role of protein phosphatases, which act to remove these regulatory phosphates from the protein kinase and its substrates. Because our analysis reveals worms to have a mature P.Tyr-signaling network, especially

REVIEW

when compared with the yeast genome, we surveyed the worm genome for protein-tyrosine phosphatases.

Our analysis reveals 83 conventional protein-tyrosine phosphatases (PTPs) plus 6 catalytic fragments and 12 additional fragments with high homology to the noncatalytic portion of other worm PTPs. In addition, there are 26 proteins classified as dual-specificity phosphatases related to MAPK phosphatases, CDC14, PRL, PIR1, CDC25, myotubularins, or PTEN lipid phosphatases. We also identify two SBF1- and one STYX-related proteins that are related to myotubularins and MAPK phosphatases yet lack the catalytic cysteine motif. These proteins are predicted to be catalytically inert yet may function as phosphoprotein-binding domains or anti-phosphatases (36). We also identify 11 inositol polyphosphate phosphatases and 65 serine-threonine phosphatases. Among the 83 PTPs, there are 57 cytoplasmic PTPs and 26 receptor-like PTPs, most of which are worm specific, lacking clear human orthologues. Exceptions include worm orthologues of the cytoplasmic PTPs; SHP2, MEG1, and MEG2, and the receptor PTPs; and PTPδ, PTPγ, PTPμ, PTPβ and IA2 (catalytically inactive). Overall, worms contain approximately the same number of tyrosine and dual-specificity protein kinases as they do tyrosine and dual-specificity protein phosphatases. This coordinate expansion in the eukaryotic lineage of both protein-tyrosine kinases and phosphatases emphasizes the biological need to maintain tight regulation of tyrosine phosphorylation. Because of the large numbers of worm-specific PTKs (FER and KIN-15 families) and worm-specific PTPs (89%, or 66 of 74), it is tempting to speculate that these unique enzymes may regulate each other's activity, or function in the same signaling pathways. Precedence for such specificity comes from genetic data indicating that the CLR-1 receptor PTP attenuates EGL-15, an FGFR orthologue, signaling in worms (37).

Conclusions. What does the worm genome sequence tell us about mammalian signal transduction? First, it has provided an ideal model to highlight the bioinformatics challenges that lie ahead with the human genome effort and allows us to test our analysis tools and database organization. Second, it lets us refine our expectations as to the diversity and absolute number of unique protein kinases that we can expect to find in the human genome. Based on our count of 493 (411 conventional and 82 PK-like proteins) worm kinases, minus the ≈197 kinases that appear to be worm-specific expansions of certain families such as the CK1, FER, and KIN-15 families, multiplied by the ≈4-fold greater number of genes in humans compared with worms, we predict the human genome to contain ≈1,100 protein kinases (PTKs and serine/threonine kinases). A similar extrapolation predicts ≈300 human protein phosphatases (PTPs, dual-specificity phosphatases, and serine-threonine phosphatases). Because our current count of human protein kinases and phosphatases stands at ≈600 and 130, respectively, we still have about half the work ahead of us. However, recent claims predict the human genome may contain as many as 140,000 genes, compared with previous estimates of ≈80,000. Such calculations would result in a significant increase in our predictions of the total number of human protein kinases and phosphatases.

We may expect to see less evolutionary expansion of protein kinases families that serve elemental cellular functions such as cell cycle control and chromosome segregation, compared with processes involved in intercellular signaling or organogenesis. However, there is already evidence for at least a 2-fold expansion in the number of CDKs and "mitotic kinases" from worms to humans. Unlike expressed sequence tag data mining and PCR-based gene discovery approaches, genomic strategies do not bias against genes whose expression is tightly regulated in a cell-, developmental-, or disease-specific manner. This point is highlighted by the identification of 650 seven-transmembrane chemoreceptors in the worm genome (1), many of which may be expressed exclusively in single neurons. Because worms have only ≈302 neurons, compared with one trillion in humans, it would not be surprising to see this selectivity in cellular expression corroborated on mining the human genome. Indeed, because many of these novel protein kinases are likely to exhibit highly restricted expression, they may prove to be excellent targets for drug discovery in the battle against human disease.

The worm serves as a much simpler and tractable organism than humans for deciphering signaling cascades. Although their P.Tyr-signaling system is quite mature—based on the content of protein-tyrosine kinases, phosphatases, and adaptor molecules—they lack much of the molecular redundancy that exists in mammals, allowing the geneticist, biochemist, and cell biologist to more readily generate an "outline" of the signaling pathways that are conserved between worms and humans. The availability of the complete worm genome provides a unique opportunity to learn about human biology. Predicted orthologous pairs of human and worm genes can be targeted by using reverse genetic approaches to identify new signaling partners or biologic functions that can then be biochemically and functionally verified in mammals.

Although worms and humans have much in common, they also have obvious differences. Worms do not have limbs or bones, or a circulatory or immune system, and they eat only bacteria. Not surprisingly, they lack several protein kinases present in humans. Over the next 2 years, we should be better able to define which protein kinases are required for these specialized functions as the genome sequences of *Drosophila* and humans are completed. Identification and classification of the proteins present in each is just a first step toward understanding the biological complexity of life.

We thank Sara Courtneidge for her input. T.H. serves on the Scientific Advisory Board of SUGEN. T.H. is a Frank and Else Schilling American Cancer Society Research Professor.

1. The *C. elegans* Sequencing Consortium (1998) *Science* **282,** 2012–2018.
2. Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T., *et al.* (1998) *Science* **282,** 2022–2028.
3. Clarke, N. D. & Berg, J. M. (1998) *Science* **282,** 2018–2022.
4. Ruvkun, G. & Hobert, O. (1998) *Science* **282,** 2033–2041.
5. Bingham, J., Plowman, G. D. & Sudarsanam, S. (1999) *J. Cell. Biochem.* in press.
6. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.* (1996) *Science* **274,** 546, 563–567.
7. Hunter, T. & Plowman, G. D. (1997) *Trends Biochem. Sci.* **22,** 18–22.
8. Garret, J. M. (1997) *Mol. Microbiol.* **4,** 8098–8120.
9. Knippschild, U., Milne, D. M., Campbell, L. E., DeMaggio, A. J., Christenson, E., Hoekstra, M. F. & Meek, D. W. (1997) *Oncogene* **15,** 1727–1736.
10. Morgan, W. R. & Greenwald, I. (1993) *Mol. Cell. Biol.* **13,** 7133–7143.
11. Morgan, W. R. (1996) *Worm Breeder's Gazette* **14,** 27.
12. Murakami, S. & Johnson, T. E. (1998) *Curr. Biol.* **8,** 1091–1094.
13. Rosato R., Veltmaat, J. M., Groffen, J. & Heisterkamp, N. (1998) *Mol. Cell. Biol.* **18,** 5762–5770.
14. Metzstein, M. M., Stanfield, G. M. & Horvitz, H. R. (1998) *Trends Genet.* **14,** 410–416.
15. Hanks, S. K., Quinn, A. M. & Hunter, T. (1988) *Science* **241,** 42–52.
16. Downward, J. (1998) *Curr. Opin. Cell Biol.* **10,** 262–267.
17. Maehama, T. & Dixon, J. E. (1999) *Trends Cell Biol.* **9,** 125–128.
18. Bellanger, J. M., Lazaro, J. B., Diriong, S., Fernandez, A., Lamb, N. & Debant, A. (1998) *Oncogene* **16,** 147–152.
19. Sun, Z., Hsiao, J., Fay, D. S. & Stern, D. F. (1998) *Science* **281,** 272–274.
20. Matsuoka, S., Huang, M. & Elledge, S. J. (1998) *Science* **282,** 1893–1897.
21. Meneghini, M. D., Ishitani, T., Carter, J. C., Hisamoto, N., Ninomiya-Tsuji, J., Thorpe, C. J., Hamill, D. R., Matsumoto, K. & Bowerman, B. (1999) *Nature (London)* **399,** 793–797.
22. Korpelainen, E. I., Kärkkäinen, M., Gunji, Y., Vikkula, M. & Alitalo, K. (1999) *Oncogene* **18,** 1–8.
23. Hou, X. S., Melnick, M. B. & Perrimon, N. (1996) *Cell* **84,** 411–419.
24. Kawata, T., Shevchenko, A., Fukuzawa, M. Jermyn, D. A., Totty, N. F., Zhukovskaya, N. V., Sterling, A. E., Mann, M. & Williams, J. G. (1997) *Cell* **89,** 909–916.
25. Williams, J. C., Wierenga, R. K. & Saraste, M. (1998) *Trends Biochem. Sci.* **23,** 179–184.
26. Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B & Kinzler, K. W. (1997) *Science* **275,** 1787–1790.
27. Leonard, C. J., Aravind, L. & Koonin, E. V. (1998) *Genome Res.* **8,** 1038–1047.
28. Brasseur, G., Tron, G., Dujardin, G., Slonimski, P. P. & Brivet-Chevillotte. P. (1997) *Eur. J. Biochem.* **246,** 103–111.
29. Hon, W.-C., McKay, G. A., Thompson, P. R., Sweet, R. M., Yang, D. S. C., Wright, G. D. & Berghuis, A. M. (1997) *Cell* **89,** 887–895.
30. Daigle, D. M., McKay, G. A., Thompson, P. R. & Wright, G. D. (1999) *Chem. Biol.* **6,** 11–18.
31. Rao, V. D., Misra, S., Boronenkov, I. V., Anderson, R. A. & Hurley, J. H. (1998) *Cell* **94,** 829–839.
32. Cote, G. P., Luo, X., Murphy, M. B. & Egelhoff, T. T. (1997) *J. Biol. Chem.* **272,** 6846–6849.
33. Ryazanov, A. G., Ward, M. D., Mendola, C. E., Pavur, K. S., Dorovkov, M. V., Wiedmann, M., Erdjument-Bromage, H., Tempst, P., Parmer, T. G., Prostko, C. R., *et al.* (1997) *Proc. Natl. Acad. Sci. USA* **94,** 4884–4889.
34. Stock, J. (1999) *Curr. Biol.* **9,** R364–R367.
35. Harris, R. A., Hawes, J. W., Popov, K. M., Zhao, Y., Shimomura, Y., Sato, J., Jaskiewicz, J. & Hurley, T. D. (1997) *Adv. Enzyme Regul.* **37,** 271–293.
36. Hunter, T. (1998) *Nat. Genet.* **18,** 303–305.
37. Kokel, M., Borland, C. Z., DeLong, L., Horvitz, H. R. & Stern, M. J. (1998) *Genes Dev.* **12,** 1425–1427.